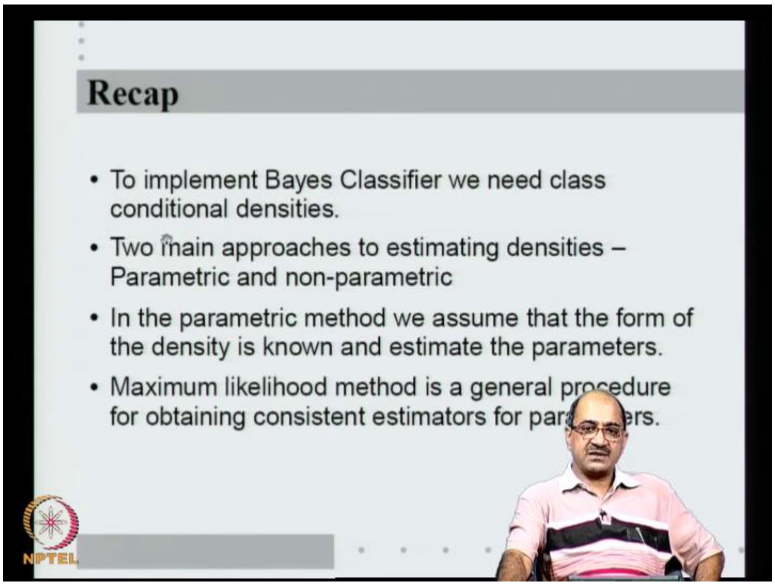
Pattern Recognition
Prof. P. S. Sastry
Department of Electronics and Communication Engineering
Indian Institute of Science, Bangalore

Lecture - 6
Maximum Likelihood Estimation of Different Densities

Welcome to the next lecture on Pattern Recognition. So, let us briefly recall what we have been doing so far. We looked at Bayes classifier and we seen that to implement Bayes classifier, we need class conditional densities. So, we have been looking at methods for obtaining class conditional densities from the training samples of examples.

(Refer Slide Time: 00:37)



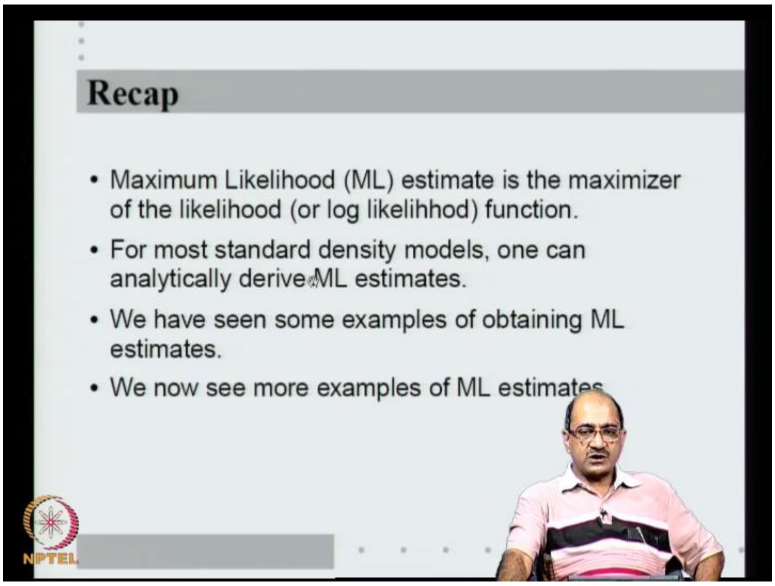
Recap

- To implement Bayes Classifier we need class conditional densities.
- Two main approaches to estimating densities – Parametric and non-parametric
- In the parametric method we assume that the form of the density is known and estimate the parameters.
- Maximum likelihood method is a general procedure for obtaining consistent estimators for parameters.

As we seen last class, there are there are essentially two approaches to estimating densities given iid examples, parametric and non-parametric. Last class, we saw that the parametric method; we assume the form of a density, that is we know the, we assume that the density is known, except for the value of some parameters. And then estimate the parameters from the training samples.

We have seen many, many ways to rate different density different estimates, such as unbiased estimates, uniformly minimum variance unbiased estimates and so on. One property we looked at is consistency of an estimator; an estimate is said to be consistent, if as sample size grows to infinity, it converges in probability to the true value of the parameter. And we have seen that maximum likelihood method is a general procedure for obtaining consistent estimators for parameters.

(Refer Slide Time: 01:35)

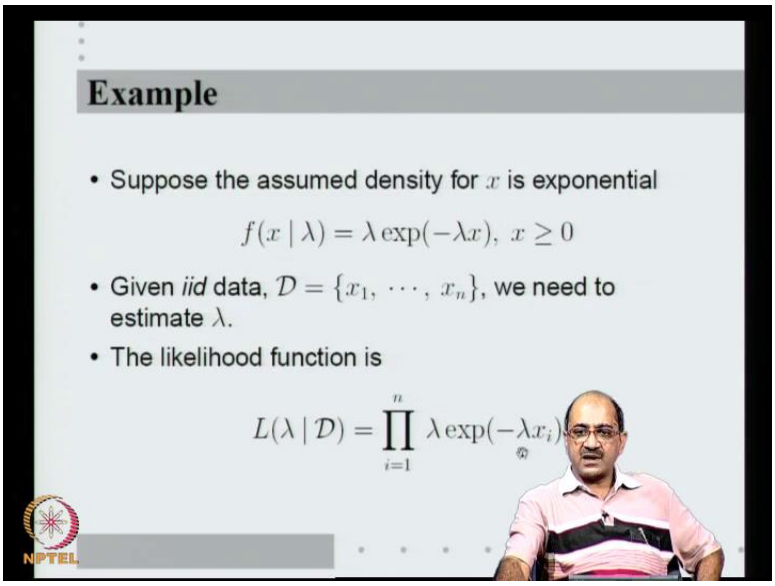


Recap

- Maximum Likelihood (ML) estimate is the maximizer of the likelihood (or log likelihhod) function.
- For most standard density models, one can analytically derive ML estimates.
- We have seen some examples of obtaining ML estimates.
- We now see more examples of ML estimates

We looked at the maximum likelihood method; in the maximum likelihood estimate, the estimate is defined at the maximizer of the likelihood function. And we have seen a few examples. But incase in most standard density models one can analytically derive the maximum likelihood estimates. Last class we seen it for a couple of examples, one is one dimensional Gaussian estimating both mean and variance. And we also seen a simple discrete random variable namely, the Bernoulli random variable that takes only two values 0 and 1 is like I toss a coin, and find the probability of heads of the coin, estimate the probability of heads of a coin. For these two simple examples, we seen out of 10 ML estimates this class, we will see a few more examples of obtaining ML estimates and then close the close the pattern maximum likelihood estimation of densities, so we will start with a couple of examples.

(Refer Slide Time: 02:35)



Example

- Suppose the assumed density for x is exponential

$$f(x \mid \lambda) = \lambda \exp(-\lambda x),\ x \geq 0$$

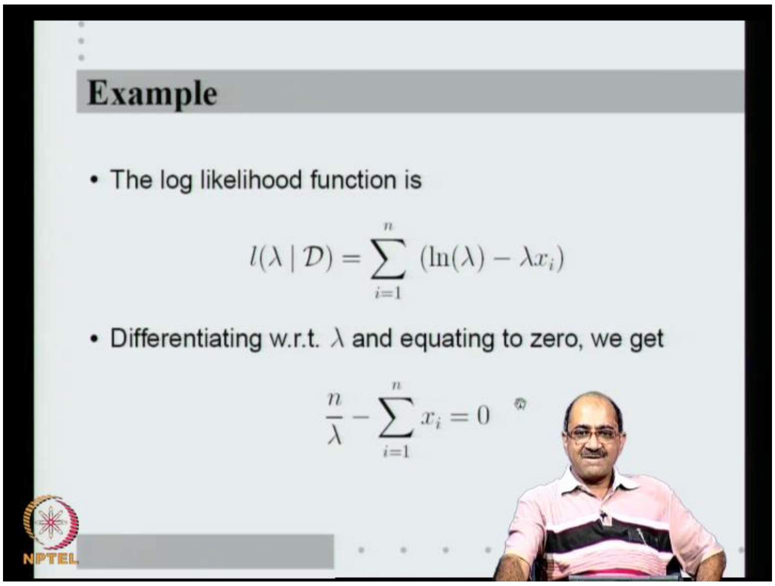
- Given *iid* data, $\mathcal{D} = \{x_1, \cdots, x_n\}$, we need to estimate λ.
- The likelihood function is

$$L(\lambda \mid \mathcal{D}) = \prod_{i=1}^{n} \lambda \exp(-\lambda x_i)$$

The first example, let us take the density to be exponential, that is our feature our feature value, which is a scalar here, has exponential density, that is that is the assumed density model. Lambda is the parameter and f x given lambda, the lambda exponential minus lambda x; this of course 0 for x less than 0, so the lambda if the parameter theta here. So, we are given iid data as, I said we can represent data either by a boldface x or by the script D, we will this class, we will use script D.

So, we are given n iid samples from this density, we do not know lambda and we need to estimate lambda given this iid examples. And for our estimation, we have to maximize the likelihood, so in this case the likelihood function is given by product i is equal to 1 to n of f x, i given lambda that is lambda exponential minus lambda x i. So, we want to find lambda, we are given x i is that is the data, we need to find lambda that maximizes this, as we said last class very often, for each of optimization we take the log likelihood. So, if I take logarithm of this, then the product becomes some I get log of lambda plus log of this will become minus lambda x i.

(Refer Slide Time: 03:56)



Example

- The log likelihood function is

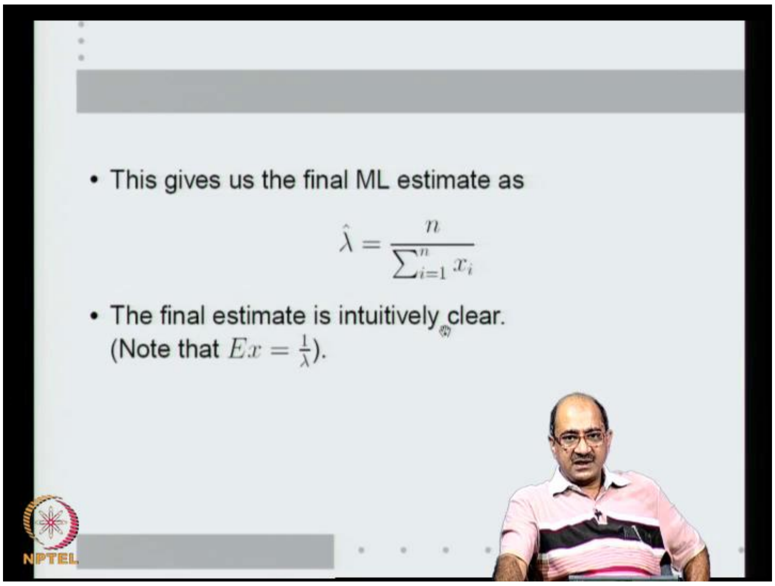
$$l(\lambda \mid \mathcal{D}) = \sum_{i=1}^{n} (\ln(\lambda) - \lambda x_i)$$

- Differentiating w.r.t. λ and equating to zero, we get

$$\frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0$$

So, the log likelihood, which we will we normally denote by small l, once again that is a function of lambda given all the data, is this product becomes sum and l n lambda and minus lambda x I, so some i is equal to one to n l n lambda minus lambda x i. So, we need to maximize this with respect to lambda right. So for that we differentiate with respect to lambda the first term will give me, 1 by lambda and there this summation, so there will be n such term so I will get an n by lambda. The second term is minus lambda summation x i, if i differentiate that with respect to lambda, i get summation x I, so differentiate with lambda and equating to 0 I get this. Now this, I can solve right I take you know n by lambda on the other side and that we can solve for lambda.

(Refer Slide Time: 04:47)



- This gives us the final ML estimate as

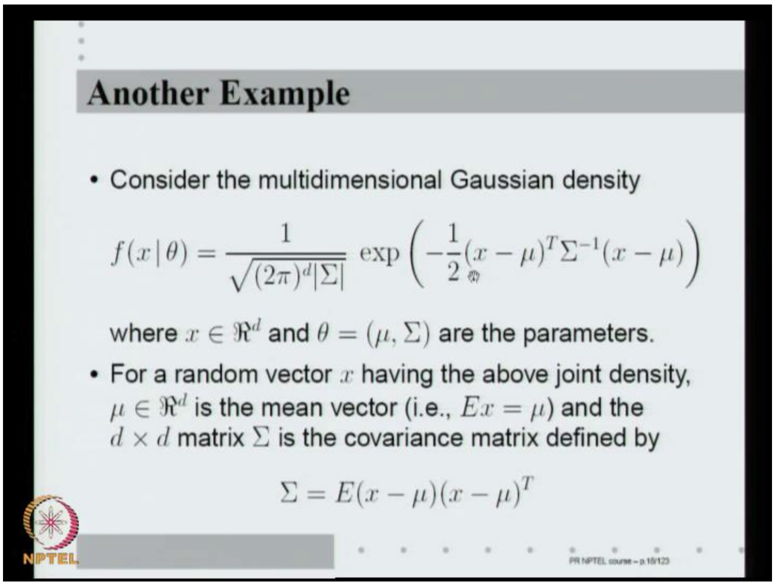
$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i}$$

- The final estimate is intuitively clear. (Note that $Ex = \frac{1}{\lambda}$).

And if, we solve our final ML estimate because, an estimate I put a hat on top is n by summation as equal to 1 to n x i right. Lambda once summation comes on that side lambda goes there and n comes down. So, the final estimate is n by i is equal to 1 to x i. This of course is is intuitively very clear because, you know for the exponential density expectation the mean is 1 by lambda right. So, because a a good estimate for mean is sample mean a good estimate for lambda is 1 by sample mean, this is simply 1 by sample mean right, 1 by n summation x i sample mean. So, n by summation x i is 1 by sample mean that is the maximum likelihood estimate for lambda.

In general if the parameters of the density are related to movements of the random variables and all movements of x are expectations. So, if we take the corresponding sample mean approximations for the expectations, very often, they turn out to be the maximum likelihood estimate for the parameters. We already seen this for example, the maximum likelihood estimate for mean of a normal random variable happens to be a sample mean. Now the maximum likelihood estimate of for lambda of a exponential density happens to be 1 by sample mean because, the actual mean is one by lambda. We will see a few more examples, so for we seen only the scalar case for continuous random variables and one scalar case for discrete random variables. So, let us look at one vector case.

(Refer Slide Time: 06:22)



Another Example

- Consider the multidimensional Gaussian density

$$f(x\,|\,\theta) = \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

where $x \in \Re^d$ and $\theta = (\mu, \Sigma)$ are the parameters.

- For a random vector x having the above joint density, $\mu \in \Re^d$ is the mean vector (i.e., $Ex = \mu$) and the $d \times d$ matrix Σ is the covariance matrix defined by

$$\Sigma = E(x-\mu)(x-\mu)^T$$

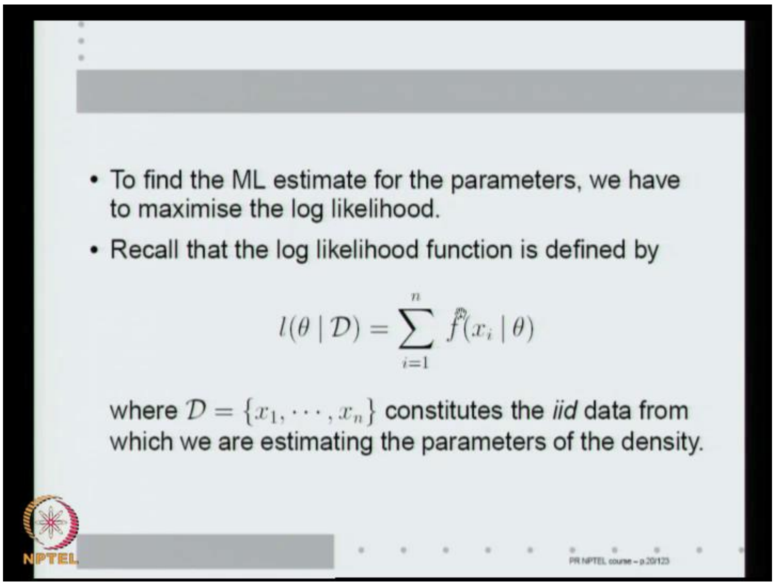
PR NPTEL course – p.18/123

So, let us look at a multidimensional Gaussian density, here x is a vector x is in d dimensional real space, x is a d component vector. This density is the d dimensional Gaussian density. So, if x is a x 1, x n, x d then joint density of all those features is given by this. I am assuming everybody has seen the multidimensional Gaussian density earlier this is 1 by root of 2 pi to the power d into determinant of sigma exponential minus of x minus mu transpose sigma inverse x minus mu, as I said our notation is that all vectors are column vectors.

So, x minus mu is a column vector, so x minus mu transpose sigma inverse x minus mu is a quadratic form which is scalar. So, here the unknown parameters theta are constituted by the mean vector mu and the matrix sigma, for this density. So, if x a vector x has the above joint density then this mu obviously, because I am taking x minus mu, mu is also in r d the d dimensional vector mu is the mean vector of this mu that is the expectation of that x will be in mu. And the sigma is a d by d matrix, which is the covariance matrix defined by this expect value of x minus mu x minus to transpose because, x minus mu is a column vector. x minus mu, x minus mu transpose is a d by d symmetric matrix.

That this is the covariance matrix and I suppose all of you know that the diagonal elements of the covariance matrix are the variances of the various components. And the off diagonal elements are the co-variances. And because this is a a covariance matrix, this is also positive definite even though, that fact does not concern as in this estimation. So, our job is given iid data from this density to estimate mu and sigma.

(Refer Slide Time: 08:13)



- To find the ML estimate for the parameters, we have to maximise the log likelihood.
- Recall that the log likelihood function is defined by

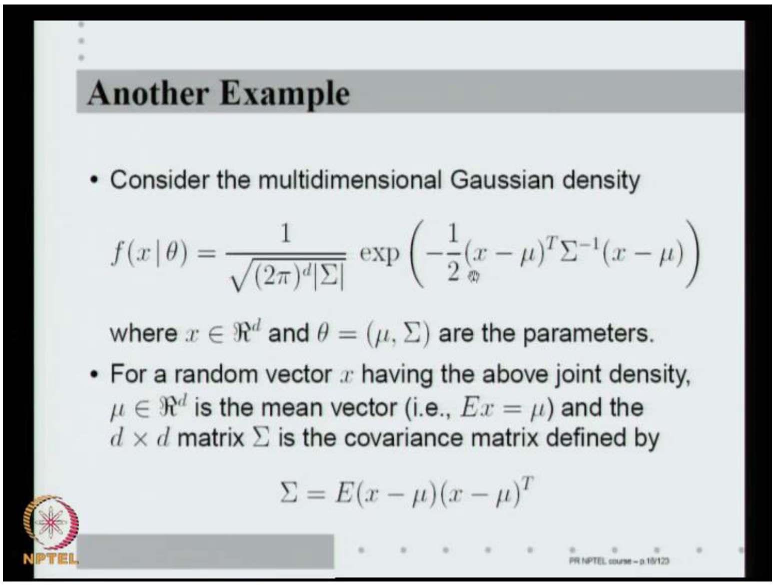
$$l(\theta \mid \mathcal{D}) = \sum_{i=1}^{n} f(x_i \mid \theta)$$

where $\mathcal{D} = \{x_1, \cdots, x_n\}$ constitutes the *iid* data from which we are estimating the parameters of the density.

PR NPTEL course – p.20/123

So, to find ML estimate, we have to maximize the log likelihood. Once again log likelihood is say likelihood is product effect say given theta. So, log likelihood is some effect say given data, where d consists of the iid sample, we have x 1 to x n.

(Refer Slide Time: 08:37)



Another Example

- Consider the multidimensional Gaussian density

$$f(x\,|\,\theta) = \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

where $x \in \Re^d$ and $\theta = (\mu, \Sigma)$ are the parameters.

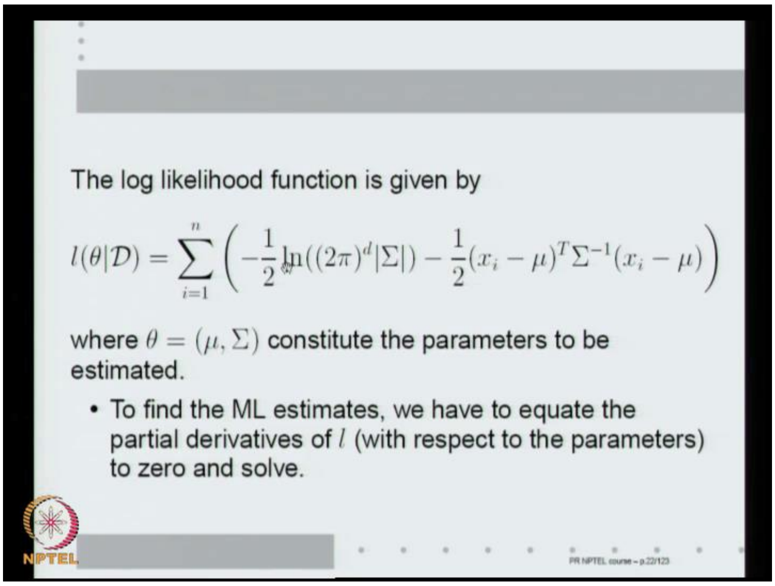
- For a random vector x having the above joint density, $\mu \in \Re^d$ is the mean vector (i.e., $Ex = \mu$) and the $d \times d$ matrix Σ is the covariance matrix defined by

$$\Sigma = E(x-\mu)(x-\mu)^T$$

PR NPTEL course – p.18/123

So, if I substitute for this effect say right into this expression i am i am sorry about it this is log. So, they will l n here the l n is missing. l n effect say, x i given theta.

(Refer Slide Time: 08:56)



The log likelihood function is given by

$$l(\theta|\mathcal{D}) = \sum_{i=1}^{n}\left(-\frac{1}{2}\ln((2\pi)^d|\Sigma|) - \frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right)$$

where $\theta = (\mu, \Sigma)$ constitute the parameters to be estimated.

- To find the ML estimates, we have to equate the partial derivatives of l (with respect to the parameters) to zero and solve.

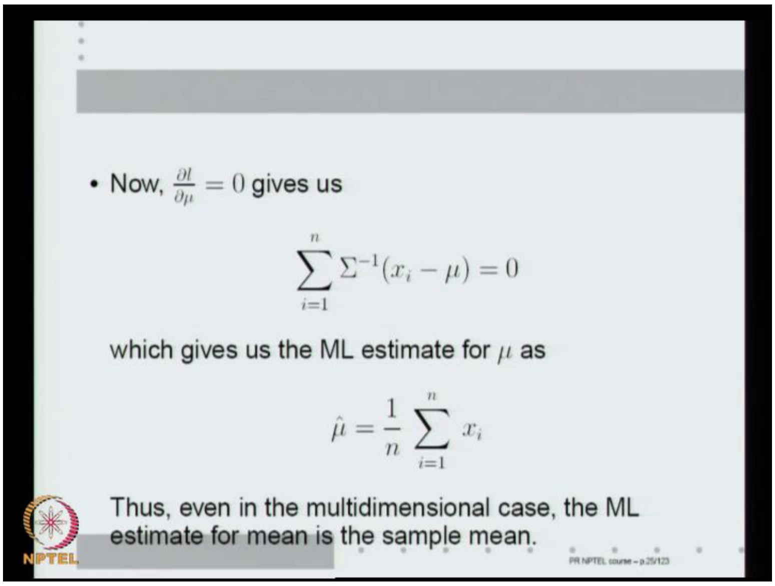
PR NPTEL course – p.22/123

So, if I take the log log likelihood, I get summation i is equal to 1 to n, I have to take log of this. So, log of this will give me, log of this term that is minus half 2 pi to power d into determinant of sigma. Log of this term will be simply minus half into that quadratic form.

So, that gives me this at the log likelihood function, summation i is equal to 1 to n minus half l n, 2 pi to power d into determinant of sigma minus half x i minus mu transpose sigma inverse x i minus mu; once again mu and sigma constitute the parameters to be estimated. So, to get this estimates, I have to differentiate the log likelihood, with respect to mu and sigma and equate at to 0 right, now mu is a vector sigma is a matrix. So, to differentiate with respect to vector that is you have to find the gradient of this function with respect to mu.

So, if I am differentiate this with respect to mu the first term is independent of mu. So, that will go to 0, the second term is a quadratic form in mu. So, it is derivative will be simply sigma inverse times x i minus mu. So, if I differentiate like that this is, what I get. If I if i take derivative with respect to the vector mu and equate to 0, what I get is summation i is equal to 1 to n sigma inverse x i minus mu is equal to 0 right. The the half does not matter because, I am equating it to 0, the quadratic forms derivative is sigma inverse x i minus mu, so that is how I get that.

(Refer Slide Time: 10:18)



- Now, $\frac{\partial l}{\partial \mu} = 0$ gives us

$$\sum_{i=1}^{n} \Sigma^{-1}(x_i - \mu) = 0$$

which gives us the ML estimate for μ as

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

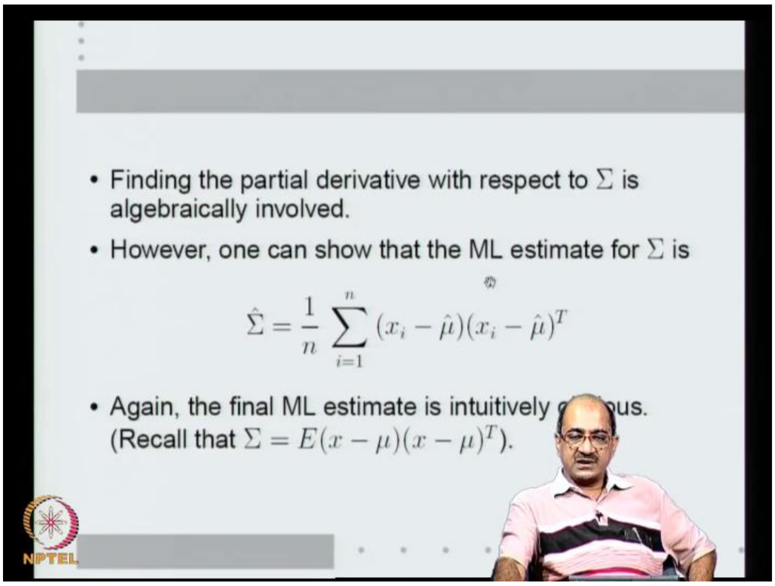
Thus, even in the multidimensional case, the ML estimate for mean is the sample mean.

PR NPTEL course – p.25/123

Now I can solve this, the the sigma matrix is non-singular right, so I can multiply both sides with sigma. So that gives me summation i is equal to 1 to n x i minus mu is equal to

0. So that gives me summation x i is equal to n times mu or mu is 1 by n summation x i. So, that gives us our final ML estimate for mu as one by n summation i is equal to one to n x I, once again, even in the vector case here, of course x i’s are vectors and mu is also a vector. So, even in the vector case that is even for the multidimensional normal distribution the ML estimate for mean, is the sample mean now what about the ML estimate for the sigma matrix. Now, that is a little more tricky because, I have to differentiate this expression with respect to a matrix, that is algebraically little more involved.

(Refer Slide Time: 11:30)



- Finding the partial derivative with respect to Σ is algebraically involved.
- However, one can show that the ML estimate for Σ is

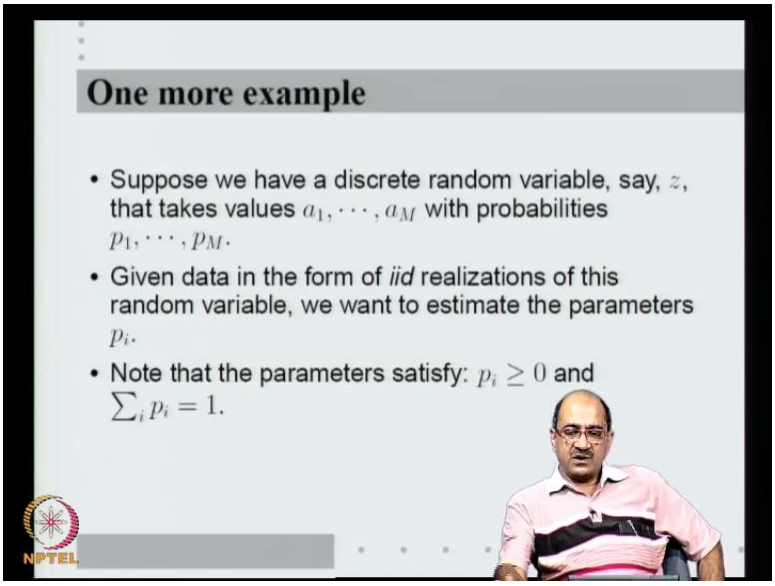
$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})(x_i - \hat{\mu})^T$$

- Again, the final ML estimate is intuitively obvious. (Recall that $\Sigma = E(x-\mu)(x-\mu)^T$).

So, we would not do the details but, one can show that, the ML estimate for sigma turns out to be 1 by n summation i is equal to 1 to n x i minus mu hat into x i minus mu hat transpose, while all the algebra escaped, I hope you can see that the final estimate is intuitively obvious. Why do I say intuitively obvious, we know how the sigma matrix is defined for the given density the sigma is nothing but, expected value of x minus mu x minus mu transpose say, we want to estimate it.

I want to take sample mean of this expectation, that will be x i minus mu x i minus mu transpose summed over i is equal to 1 to n divided by 1 by n but, I do not know mu, so I have chose mu hat. So, intuitively this is what we should expect the ML estimate for sigma to be and it turns to be that even though because, the algebra is rather involved or just skip the algebra.

(Refer Slide Time: 12:25)



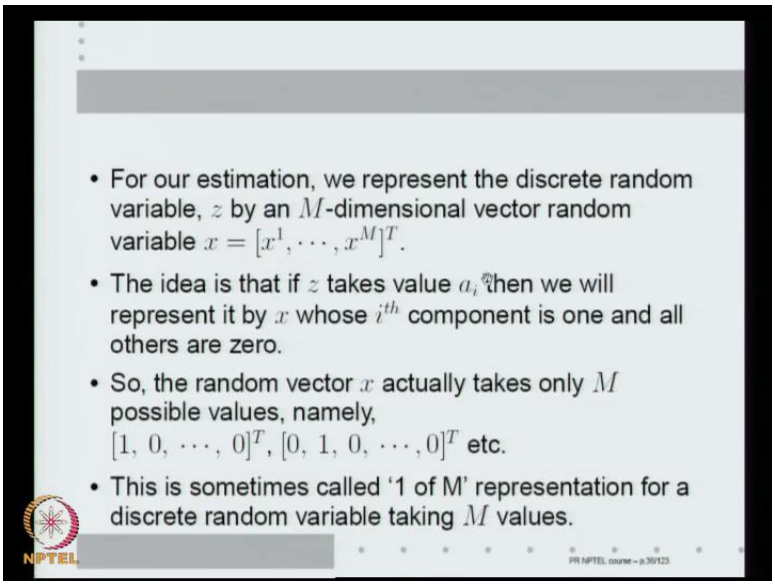
One more example

- Suppose we have a discrete random variable, say, z, that takes values $a_1, \cdots, a_M$ with probabilities $p_1, \cdots, p_M$.
- Given data in the form of *iid* realizations of this random variable, we want to estimate the parameters p_i.
- Note that the parameters satisfy: $p_i \geq 0$ and $\sum_i p_i = 1$.

So, let us look at one more example, before we finish m l estimate. This time let us look at a discrete random variable again. We looked at a few continuous random variables. So, suppose we have a discrete random variables z, now this time let us not take a specific discrete random variable any generic discrete random variable. Every discrete random variable takes let us say is the discrete random variable taking only finitely many values. Let us say a 1 to a m are the values that the random variable takes and the corresponding probability say p 1 to p m.

So, what is that we have to do, we are given data in the form of iid realizations of this z. And we want to estimate the parameters p i. ofcourse, The parameters p i have to satisfy p i defined the mass probability mass function of z p i's are probabilities. So, pi is have to be greater than or equal to 0 and summation p i is equal to 1 because, summation of the mass function should be equal to 1. So, the parameters have to satisfy this extra constraints.

(Refer Slide Time: 13:25)



- For our estimation, we represent the discrete random variable, z by an M-dimensional vector random variable $x = [x^1, \cdots, x^M]^T$.
- The idea is that if z takes value a_i then we will represent it by x whose i^{th} component is one and all others are zero.
- So, the random vector x actually takes only M possible values, namely, $[1, 0, \cdots, 0]^T$, $[0, 1, 0, \cdots, 0]^T$ etc.
- This is sometimes called '1 of M' representation for a discrete random variable taking M values.

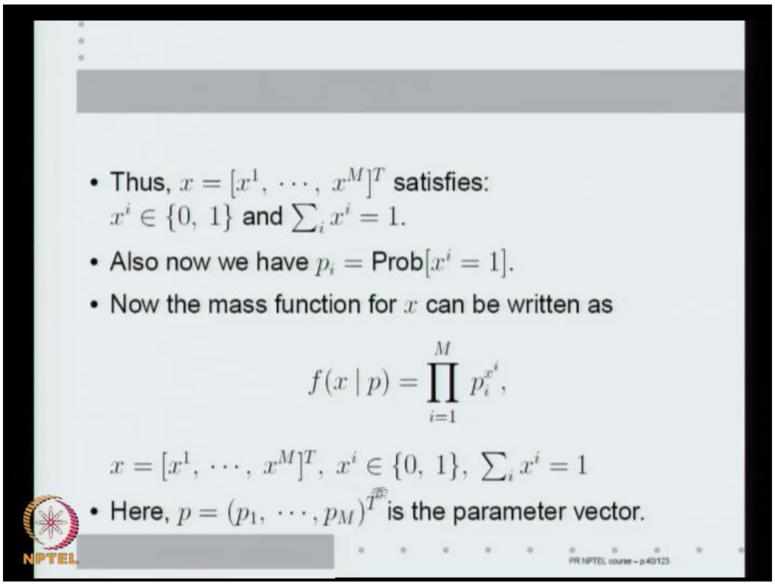
PR NPTEL course – p.39/123

Now, for deriving our ML estimate, we will, we represent our random variable in a slightly different style. So, I will represent our discrete random variable z as an n dimensional vector random variable x. Instead of thinking of if it is a scalar taking m different values, we think of it as a vector of M components, what is the idea. The idea is that if z takes the ith value say, a i then we will represent it by x whose ith component is 1 and all others are 0. So, a z that takes value a i is represented as a, so, to say ith coordinate vector in the m dimensional space.

So, what it means is that the random variable x actually takes only m possible values. Obviously, because z can take only M possible values and the values are 1, 0, 0, 0 or 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0 and so on right, each of the coordinate vectors are the only possible values for this vector x. So, we are going to represent values of z by this this is nice because essentially in estimating p 1 to p or p m the actual values a 1, a 2, a m do not make any difference to us right.

The values assumed by z are a 1, a 2, a m; so if a is a different the actual data as we see it will be different, but the specific numerical values of a 1, a 2 do not make any difference to the probabilities, the probability say p 1 is the probability of z taking, it is first value say a 1. Hence this kind of representation allows us to look at Ml estimate a parameters of the discrete random variables. Any Discrete Random Variables in a Uniform Framework, it is often, called 1 of m representation for discrete random variable taking M values. So, there are few things about this representation now, we we are thinking of a random variable as a m component vector.

(Refer Slide Time: 15:22)



- Thus, $x = [x^1, \cdots, x^M]^T$ satisfies: $x^i \in \{0, 1\}$ and $\sum_i x^i = 1$.
- Also now we have $p_i =$ Prob$[x^i = 1]$.
- Now the mass function for x can be written as

$$f(x \mid p) = \prod_{i=1}^{M} p_i^{x^i},$$

$x = [x^1, \cdots, x^M]^T, \; x^i \in \{0, 1\}, \; \sum_i x^i = 1$

- Here, $p = (p_1, \cdots, p_M)^T$ is the parameter vector.

PR NPTEL course – p.40/123

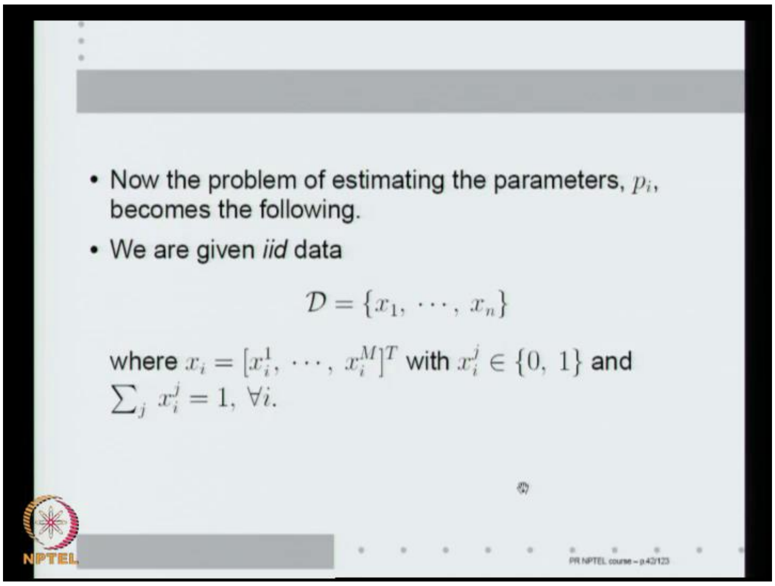
And this representation satisfies the following, so each each value of z is a value of x. So, x is a m vector but, the m vector satisfies each component of the m vector is either 0 or 1 and some of the components is 1. So, these two together tell you that exactly 1 component is 1 and all others are 0. So, this is what our representation satisfies, also probability x i is equal to 1 is given by p i, when will the x superscript i be 1 x, superscript i is the ith component of x ith component of x is 1; only if z has taken the ith value namely a i and the probability with if z takes that value is p i.

Just one classification you may be wondering why I am representing the components of a vector as superscripts. Normally when we take components of vector, who are represent of subscript I cannot represent them as subscripts because, we already agreed that our data is x 1, x 2, x n. Where x subscript 1 x subscript 2 is the representation for different data items, irrespective whether x is Scalar or Vector. So, even for this vector x the different iid data will be represented as x 1, x 2. And hence for components, we are representing it as superscript.

So, given this now I can i can write the mass function of this vector random variable x as follows, f of x given p p is my parameter vector consisting of p 1, p 2, p m is product i is equal to 1 to m p i to the power x i, because x takes only these values, x can take the possible values of x or x as an m component vector the components being x superscript one x superscript m. Each of the component say is a 0 or 1 and some of the components is one meaning exactly one component is 1 all others are 0.

So, if x for example, takes value 1, 0, 0, 0 then f of x given p will be p 1 to the power 1 and all others to the power 0, so this is p 1. So, this is exactly what is probability, this is a mass function. So, probability x is equal to 1, 0, 0, 0 is p 1 probability x is equal to 0, 1, 0, 0 is p 2 and so on. So, this is the right expression for the mass function of x, so given this mass function of x, and iid samples of x, we need to estimate the parameter vector p 1 to p m.

(Refer Slide Time: 17:56)



- Now the problem of estimating the parameters, p_i, becomes the following.
- We are given *iid* data

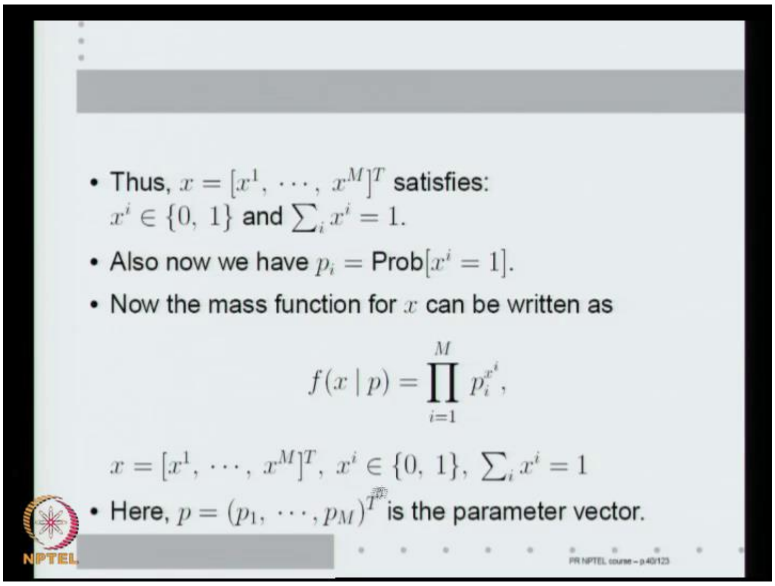
$$\mathcal{D} = \{x_1, \cdots, x_n\}$$

where $x_i = [x_i^1, \cdots, x_i^M]^T$ with $x_i^j \in \{0, 1\}$ and $\sum_j x_i^j = 1, \forall i$.

PR NPTEL course – p.43/123

So, what is the problem now, for the problem for estimating the parameters p i is you are given iid data of x’s. So, whatever z data, we have we can recode them into the corresponding samples x, knows that each x i here is a vector right. So, each x i has m component, so the first component of x i is x i superscript 1, the second component is x i superscript 2 and so on. So, x i is an m vector with, each component being either 0 or 1 and for each i summed over j x i superscript j is equal to 1 for all i. So, given data like this this is the iid realizations of z on which we are estimating p's. We know the probability mass function of x that is what, we written in written just now.

(Refer Slide Time: 18:49)



- Thus, $x = [x^1, \cdots, x^M]^T$ satisfies: $x^i \in \{0, 1\}$ and $\sum_i x^i = 1$.
- Also now we have $p_i = \text{Prob}[x^i = 1]$.
- Now the mass function for x can be written as

$$f(x \mid p) = \prod_{i=1}^{M} p_i^{x^i},$$

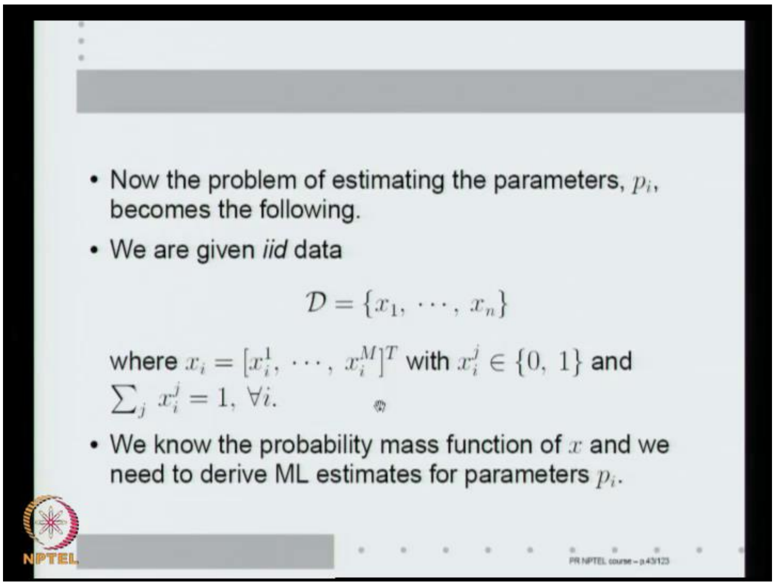
$x = [x^1, \cdots, x^M]^T$, $x^i \in \{0, 1\}$, $\sum_i x^i = 1$

- Here, $p = (p_1, \cdots, p_M)^T$ is the parameter vector.

PR NPTEL course – p.40/123

This is the probability mass function of x, the parameter the parameter p

(Refer Slide Time: 18:55)



- Now the problem of estimating the parameters, p_i, becomes the following.
- We are given *iid* data

$$\mathcal{D} = \{x_1, \cdots, x_n\}$$

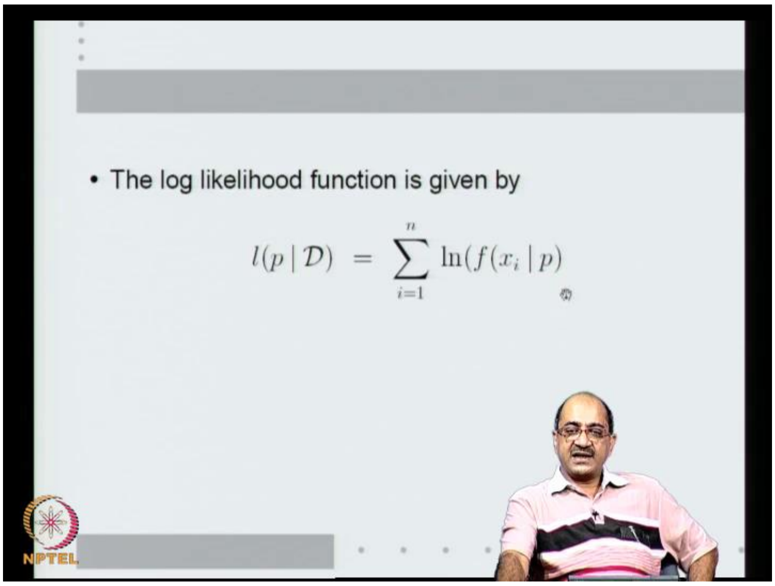
where $x_i = [x_i^1, \cdots, x_i^M]^T$ with $x_i^j \in \{0, 1\}$ and $\sum_j x_i^j = 1, \forall i$.

- We know the probability mass function of x and we need to derive ML estimates for parameters p_i.

PR NPTEL course – p.43/123

So, even the probability mass function of x, we need to derive the ML estimates for the parameters p i.

(Refer Slide Time: 19:02)

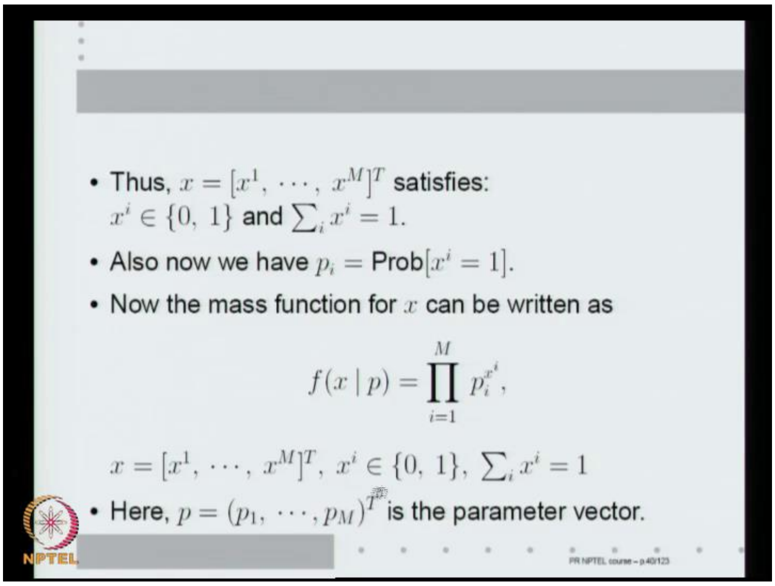


- The log likelihood function is given by

$$l(p \mid \mathcal{D}) = \sum_{i=1}^{n} \ln(f(x_i \mid p)$$

So, what we have to do log likelihood function, so what is my log likelihood function, l p given data is i is equal to 1 to n, l of f of x i given p.

(Refer Slide Time: 19:20)



- Thus, $x = [x^1, \cdots, x^M]^T$ satisfies: $x^i \in \{0, 1\}$ and $\sum_i x^i = 1$.
- Also now we have $p_i = \text{Prob}[x^i = 1]$.
- Now the mass function for x can be written as

$$f(x \mid p) = \prod_{i=1}^{M} p_i^{x^i},$$

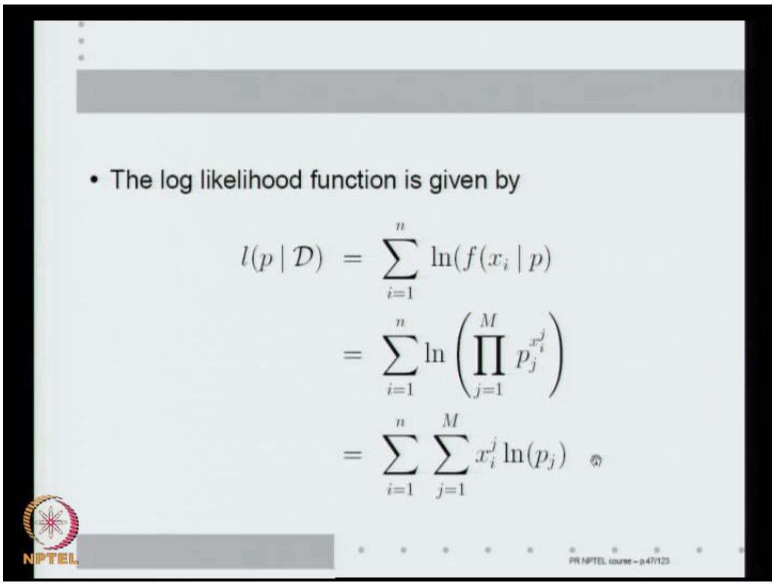
$x = [x^1, \cdots, x^M]^T$, $x^i \in \{0, 1\}$, $\sum_i x^i = 1$

- Here, $p = (p_1, \cdots, p_M)^T$ is the parameter vector.

PR NPTEL course – p.40/123

Now, x i given p is given by this right, so let us substitute that.

(Refer Slide Time: 19:24)



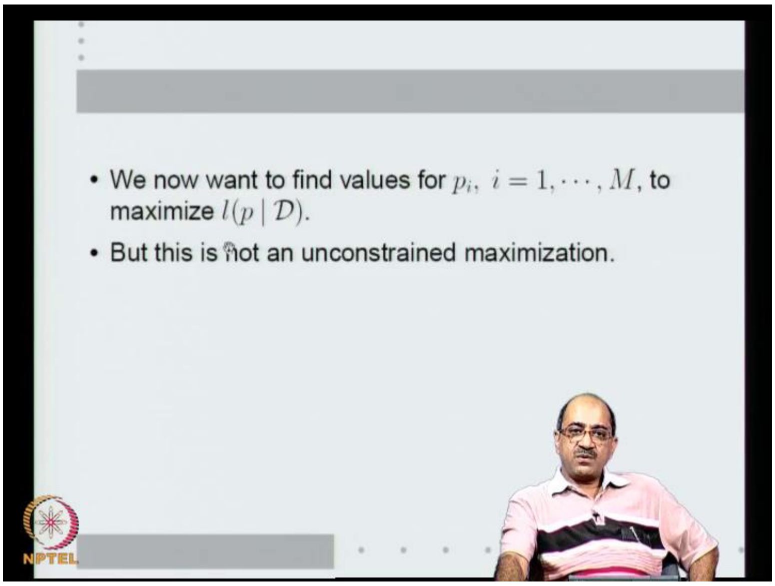
- The log likelihood function is given by

$$\begin{aligned} l(p \mid \mathcal{D}) &= \sum_{i=1}^{n} \ln(f(x_i \mid p)) \\ &= \sum_{i=1}^{n} \ln\left(\prod_{j=1}^{M} p_j^{x_i^j}\right) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{M} x_i^j \ln(p_j) \end{aligned}$$

PR NPTEL course – p.47/123

So, that is l n of x i given p is product j is equal to 1 to m p j x i j. Now because, of the l n the inner product also becomes sum. So, my log likelihood now becomes i is equal to 1 to n j is equal to 1 to m x i j l n p j. So, this is my log likelihood as a function of p and this is what i want to maximize to find my p's.

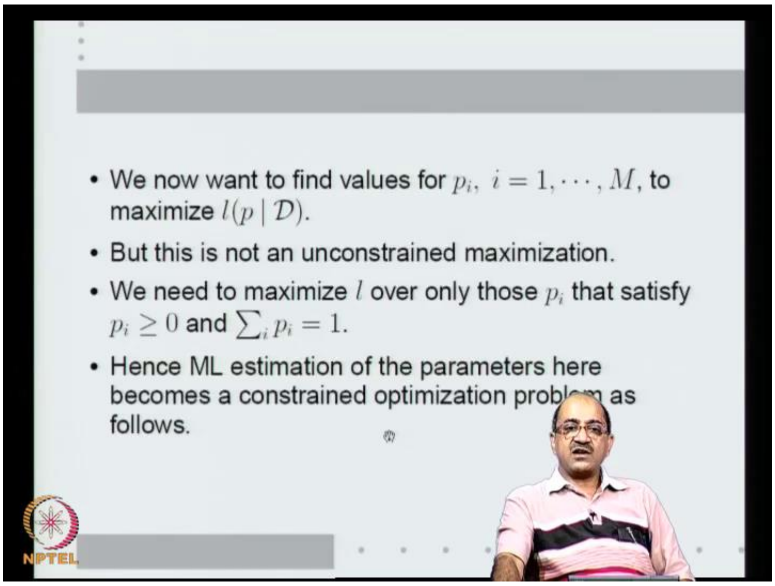
(Refer Slide Time: 19:50)



- We now want to find values for p_i, $i = 1, \cdots, M$, to maximize $l(p \mid \mathcal{D})$.
- But this is not an unconstrained maximization.

So, we want to find the parameters, so values for p i to maximize the log likelihood l p given d. But we should understand that is not an ordinary maximization problem is not an unconstrained maximization. By unconstrained maximization, I mean I do not want to maximize this over all possible m triples of real numbers p 1, p 2, p m right. Why because, we need to maximize my log likelihood over only those m triples or numbers p i that satisfy this constraint because, this p i have to be a distribution.

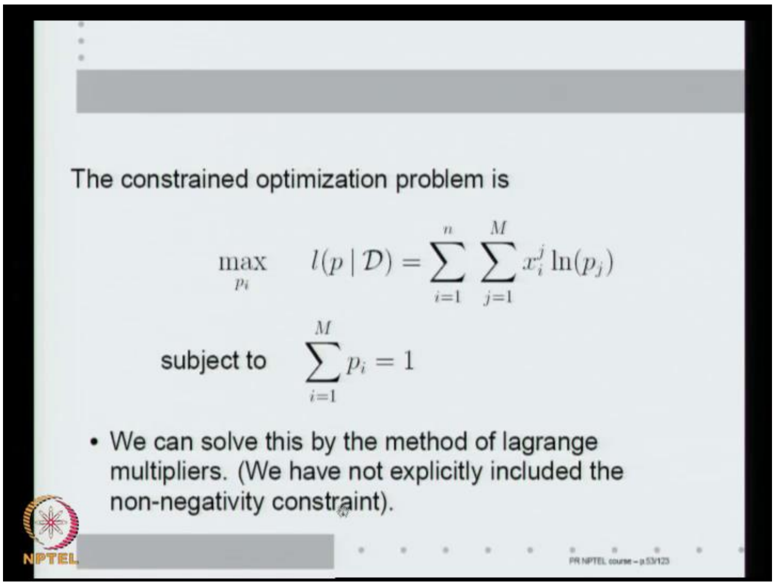
(Refer Slide Time: 20:31)



- We now want to find values for $p_i, \ i = 1, \cdots, M$, to maximize $l(p \mid \mathcal{D})$.
- But this is not an unconstrained maximization.
- We need to maximize l over only those p_i that satisfy $p_i \geq 0$ and $\sum_i p_i = 1$.
- Hence ML estimation of the parameters here becomes a constrained optimization problem as follows.

So, the maximization of log likelihood now is not an unconstrained maximization problem but, it is a constraint maximization problem. So, the ML estimate now becomes a constraint optimization problem.

(Refer Slide Time: 20:38)



The constrained optimization problem is

$$\max_{p_i} \quad l(p \mid \mathcal{D}) = \sum_{i=1}^{n} \sum_{j=1}^{M} x_i^j \ln(p_j)$$

subject to $\sum_{i=1}^{M} p_i = 1$

- We can solve this by the method of lagrange multipliers. (We have not explicitly included the non-negativity constraint).

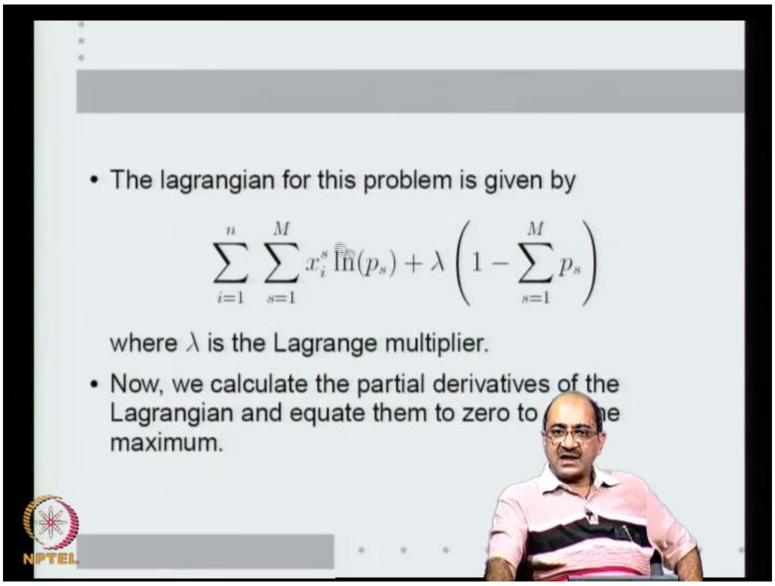
PR NPTEL course – p.53/123

What is the constraint optimization problem now, We want to maximize l p given d, which is given by, this expression subject to summation i is equal to 1 to M p i is equal to 1. So, I hope all of you know about constraint optimization problems, so constraint optimization comes as not simply maximize some function but, maximize some function subject to something and because, what we have is an equality constraint. We can do it using what is called Lagrange multipliers.

So, we are going to solve this Lagrange multipliers there is just one point, I want to draw your attention to. When we said this becomes a constraint optimization, we said p has to satisfy two constraints, p i greater than equal to 0 and summation, p i is equal to 1. But, when I formulated the constraint optimization problem, I have not explicitly included the non negativity constraint the reason is that it is easy to solve the constraint optimization problem. If I have only equality constraint then Lagrange multiplier method always gives me exact analytical solution much more easily.

So, what we can do is we can incorporate only this constraint find what is the maximum, if that also satisfies non negativity. We are done, then we even if we include the constraint, would have would have still have got the same solution. So, let us only put the summation i p i equal to one constraint and then solve it. And then we will verify that what, we get also satisfies the non negativity constraint.

(Refer Slide Time: 22:00)



- The lagrangian for this problem is given by

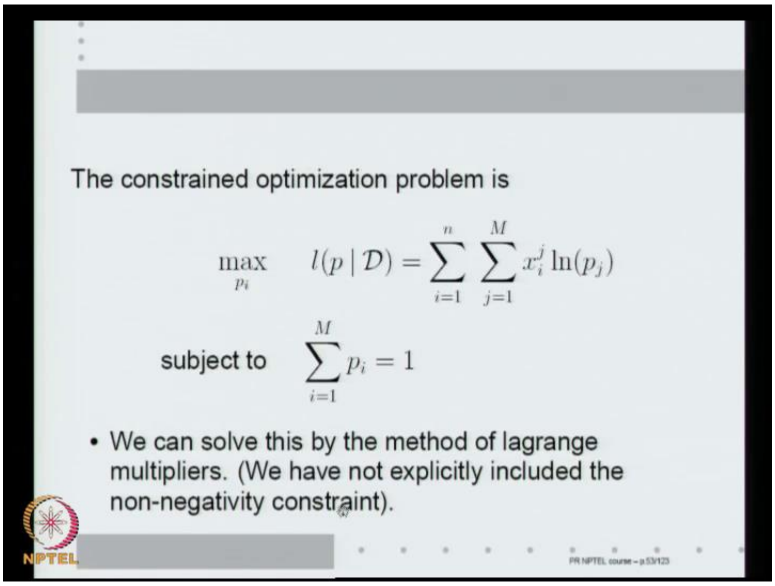
$$\sum_{i=1}^{n}\sum_{s=1}^{M} x_i^s \ln(p_s) + \lambda\left(1-\sum_{s=1}^{M} p_s\right)$$

where λ is the Lagrange multiplier.

- Now, we calculate the partial derivatives of the Lagrangian and equate them to zero to [illegible] the maximum.

So, how do I solve the constraint optimization problem, haw to find the Lagrangian.

(Refer Slide Time: 22:05)



The constrained optimization problem is

$$\max_{p_i} \quad l(p \mid \mathcal{D}) = \sum_{i=1}^{n}\sum_{j=1}^{M} x_i^j \ln(p_j)$$

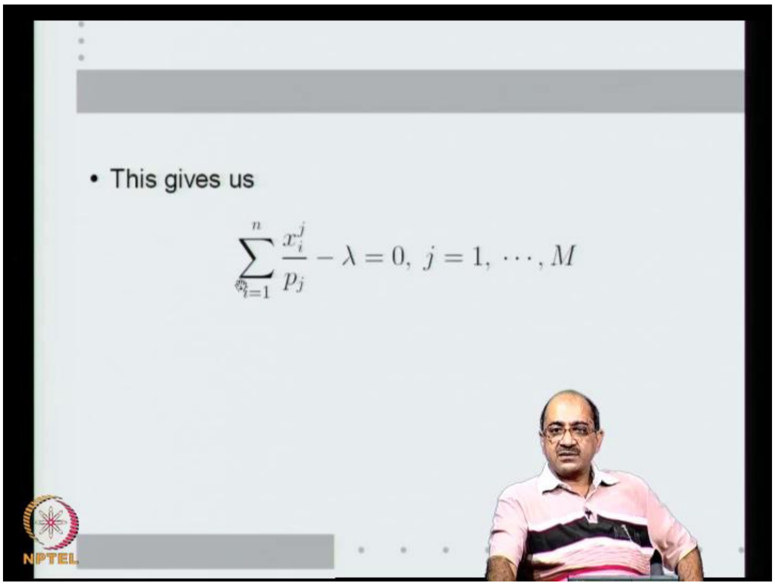
subject to $\displaystyle\sum_{i=1}^{M} p_i = 1$

- We can solve this by the method of lagrange multipliers. (We have not explicitly included the non-negativity constraint).

PR NPTEL course – p.53/123

So, constraint optimization problems comes as, maximize some objective function subject to some constraints. So, you form the Lagrangian by adding the constraint function, which here happens to be summation i is equal to 1 to m p i minus 1, we have to add that to the objective function right with a Lagrange multiplier. So, that becomes the Lagrangian for this problem. This is the objective function and this is, what we are trying to maximize; and this is the constraint function. 1 minus summation p s equal to 0 and lambda is the Lagrange multiplier. I have just changed the summation index from s to j because, we have been thinking of p j's as the parameter. Because this j and i are dummy indices they are indices of the summation.

(Refer Slide Time: 23:43)

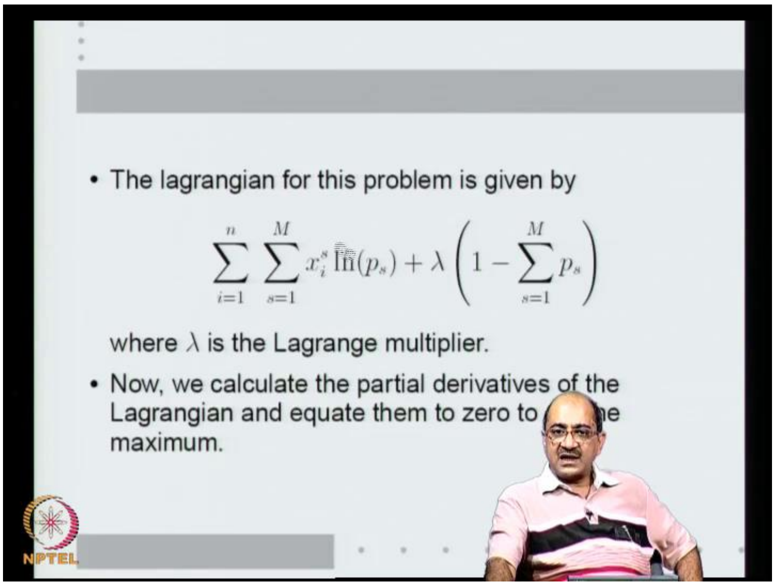


- This gives us

$$\sum_{i=1}^{n} \frac{x_i^j}{p_j} - \lambda = 0, \ j = 1, \cdots, M$$

So, my first term in the derivative, will be x i j by p j summation i is equal to 1 to n.

(Refer Slide Time: 22:52)



- The lagrangian for this problem is given by

$$\sum_{i=1}^{n}\sum_{s=1}^{M} x_i^s \ln(p_s) + \lambda\left(1 - \sum_{s=1}^{M} p_s\right)$$

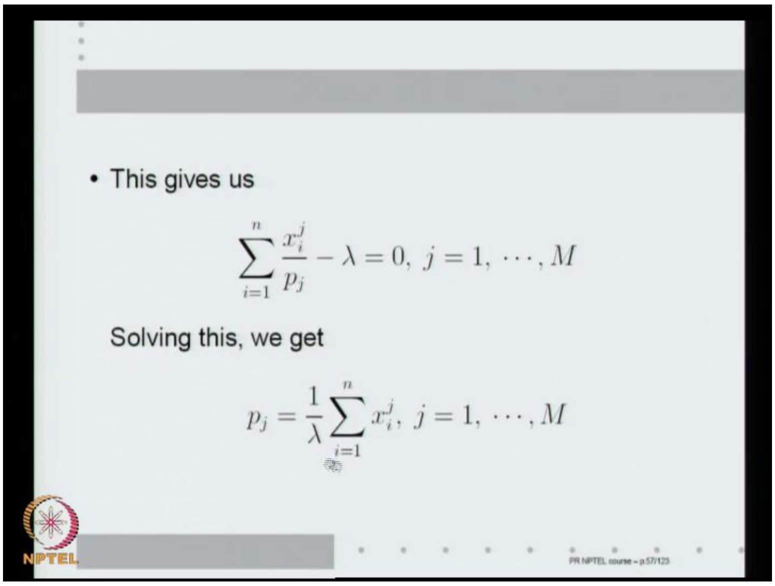
where λ is the Lagrange multiplier.

- Now, we calculate the partial derivatives of the Lagrangian and equate them to zero to [illegible]e maximum.

So, that while differentiating we would not get confused, I have just changed the dummy index j to s in this expression. Now, what do we have to do I have to take partial derivatives of this expression and equate it to 0, say let us say I take partial derivative with respect to p j for some j. So, the derivatives goes inside the first summation i is equal to 1 to n in the second summation only when s is equal to j is this expression a function of p j. Otherwise it is not a function of p subscript j.

So, when I take the derivative of this the the first term with respect to p j it becomes, i is equal to 1 to n and the derivative of this when s is equal to j that will be x i j by p j right. What is the the second term derivative, I have to differentiate this with respect to p j this is a constraint. So, this is lambda times p 1 plus lambda times p 2 and so on. So, i differentiate with respect to p j i get lambda.

(Refer Slide Time: 24:17)



- This gives us

$$\sum_{i=1}^{n} \frac{x_i^j}{p_j} - \lambda = 0, \ j = 1, \cdots, M$$

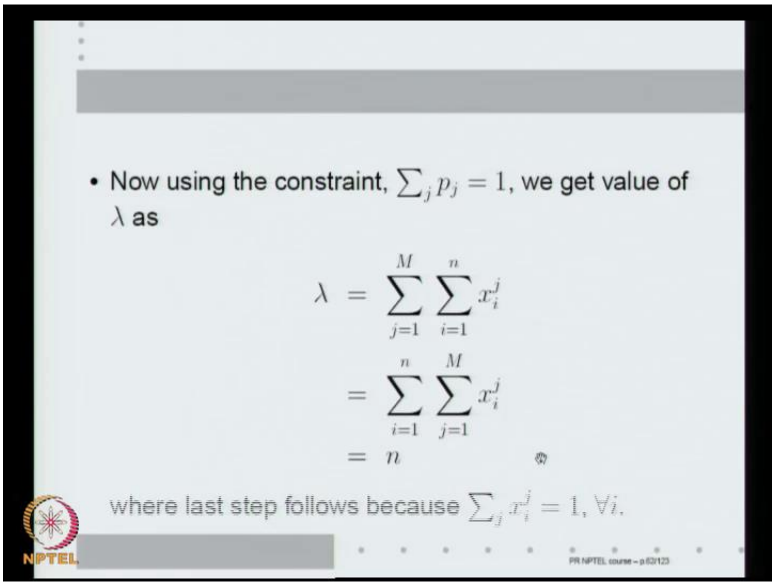
Solving this, we get

$$p_j = \frac{1}{\lambda} \sum_{i=1}^{n} x_i^j, \ j = 1, \cdots, M$$

PR NPTEL course – p.57/123

So, I equate the derivative to 0. I get summation i is equal to 1 to n x i j by p j minus lambda equal to 0 j is equal to 1 to m. So, these are the all the j differentiating with respect to all all the j's. So, I can solve this to get my p j's as 1 by lambda summation i is equal to 1 to n x i j ofcourse. Lambda is a Lagrange multiplier I still do not know it is value, I have to find the value Lagrange multiplier to complete my estimate. So, If i sum this up summation over j one by lambda summation, i is equal to 1 to n x i j should be equal to 1 and they can take lambda on that that side. The constraint is you sum this expression over j that becomes 1. So, then I am taking lambda the other side to get this.

(Refer Slide Time: 24:31)



- Now using the constraint, $\sum_j p_j = 1$, we get value of λ as

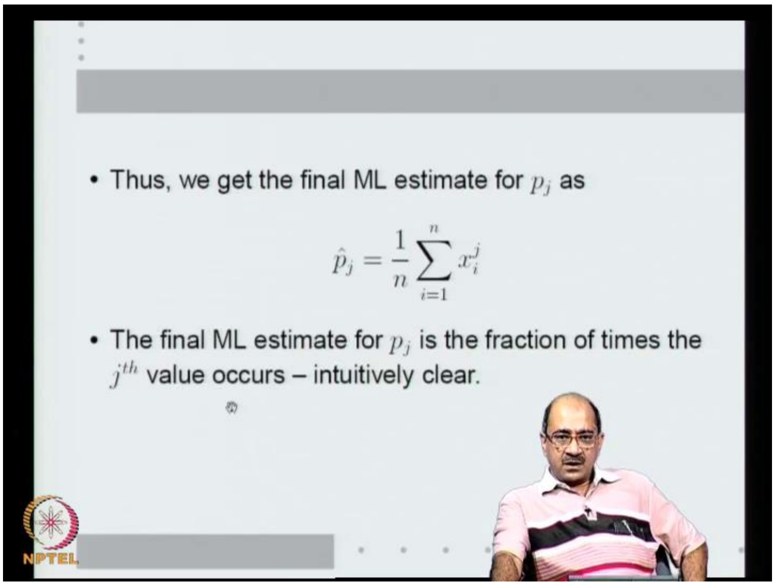
$$\begin{aligned}\lambda &= \sum_{j=1}^{M} \sum_{i=1}^{n} x_i^j \\ &= \sum_{i=1}^{n} \sum_{j=1}^{M} x_i^j \\ &= n\end{aligned}$$

where last step follows because $\sum_j x_i^j = 1, \forall i$.

PR NPTEL course – p.63/123

So, how do I get the value Lagrange multiplier by using my equality constraint, I have summation j p j is equal to 1 as my constraint. Now this is a double summation both are finite summations. So, i can change the order of the summation. So, first sum with respect to i then sum with respect to j now that becomes equal to n why this inner summation as you seen our x i j’s are such for any in any vector x i exactly one component is one all others are 0 right. So, this summation always gives me one and because of the auto summation i get n. So, the last step follows because summation j is equal to one to m x i j is one for all i. So, this is the value of lambda if. So, that constraint gives me lambda is equal to summation over j is equal to 1 to m summation of i as equal to 1 to n x i j i hope this is clear.

(Refer Slide Time: 25:45)



- Thus, we get the final ML estimate for p_j as

$$\hat{p}_j = \frac{1}{n}\sum_{i=1}^{n} x_i^j$$

- The final ML estimate for p_j is the fraction of times the j^{th} value occurs – intuitively clear.

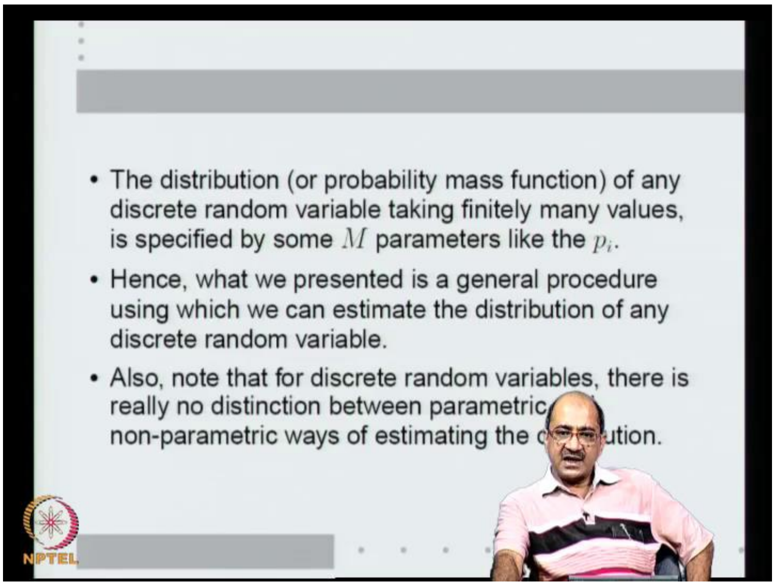
Now, plug in the value of lambda I get my final m l estimate for p j as 1 by n summation, i as equal to 1 x i j here this is been quite an optimization problem. We have actually formulated the constraint optimization problem used Lagrange multipliers and solve it. But, at the end what we get is very very intuitively obvious expression, p j is the probability with which the random variable takes jth value right.

If I sum i is equal to 1 to n x i j that sum will be exactly the number of time the jth value is taken jth value is taken x i j is 1, otherwise it is 0. So, the final estimate is the fraction of times the jth value occurs. The probability of the random variable taken the jth value is simply equal to the fraction of times the jth value occurs right it is no more than. If I want to estimate the probability of heads of a coin then a good estimate is the fraction of heads out of n tosses right.

As a matter of fact, we have seen that that density in the last class the Bernoulli density, that is a that is a discrete random variable taking only two values 0 and 1. This is actually generalization this is a discrete random variable taking m different values right. So, If I take m is equal to two this gives me the same thing at the Bernoulli random variable.We considered last class only thing is when, when m is equal to 2, I do not need to keep p 1 and p 2 because p 1 plus p 2 is equal to 1 any one of them will do.

So, we will take probability of the first value as the parameter, when you have m we keep the constraint that summation p i is equal to 1 and estimate all the p i. So, in that sense this is a generalization of the binary Bernoulli random variables we had taken.

(Refer Slide Time: 27:30)



- The distribution (or probability mass function) of any discrete random variable taking finitely many values, is specified by some M parameters like the p_i.
- Hence, what we presented is a general procedure using which we can estimate the distribution of any discrete random variable.
- Also, note that for discrete random variables, there is really no distinction between parametric non-parametric ways of estimating the d ution.

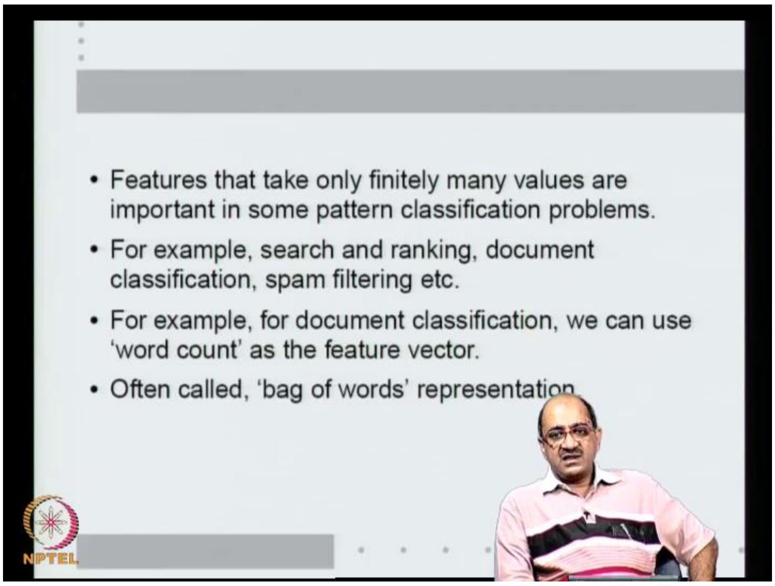
So, let us sum up this example this example is not just an example. It is its actually much more than a simple example. Because, the distribution that is the probability mass function of any discrete random variable that takes finitely many values is simply specified by some m numbers p i right. The numbers have to satisfy p i greater than or equal to 0 or summation p i is equal to 1 except for that every discrete random variable mass function is simply specified by this parameters p i.

Hence what we have just now got is a general procedure for estimating the distribution of any discrete random variable right; not one specific discrete random variable like Bernoulli here. Binomial or Poisson or any Poisson ofcourse, takes infinitely many values. But, this is a general procedure for estimating the distribution of any discrete random variable also when we consider discrete random variables there is really no distinction between parametric and non parametric ways of estimating.

Because a discrete random variable taking finitely many value the the most general distribution is still specified by m numbers. So, when I say, I am estimating the m numbers. I am not constraining the random variables distribution in any way right. In the continuous case, when I am assuming the density to be normal or exponential is an assumption is a distinct to assumption, the real data may or may not satisfy the assumption of the distribution. That I have assumed on the other hand when, I am when i am estimating the distribution of a discrete random variable taking finitely.

Many values this procedure, we just now laid out simply. I assume that the distribution is specified by m numbers p1, p 2, p m all of them are non negative. And they sum to 1 and that does not constraint the mass function as a discrete random variable in anyway which means, what we have here is a way to estimate the probability mass function of any discrete random variable if you are given iid samples.

(Refer Slide Time: 29:33)



- Features that take only finitely many values are important in some pattern classification problems.
- For example, search and ranking, document classification, spam filtering etc.
- For example, for document classification, we can use 'word count' as the feature vector.
- Often called, 'bag of words' representation

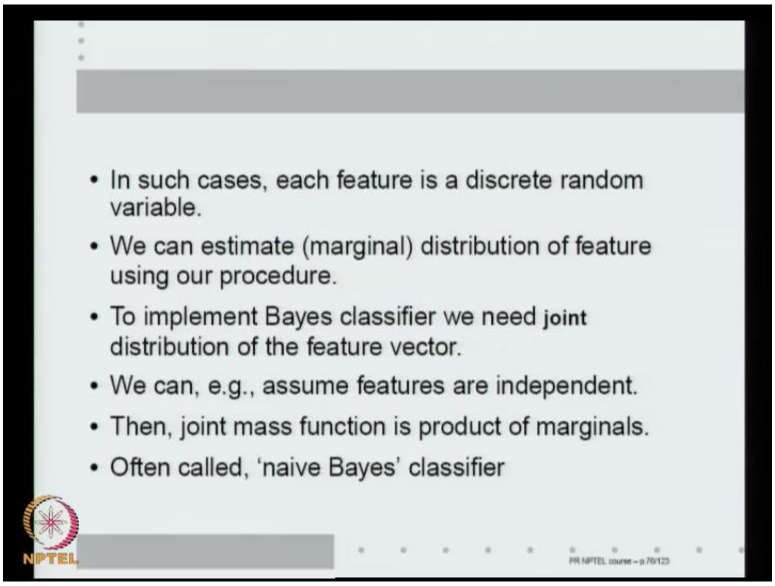
Now, discrete random variables are also important in many pattern recognition problems. Even though many pattern recognition problems also come with continuous random variables. But, discrete random variable taking finitely many values or particularly important is some of the recent problems which are, mostly web based search and ranking, document classifications, spam filtering and so on right.

For example, If I want to classify HTML pages as advertisement pages or information pages, ultimately the features. I measure is the number of links it has kinds of words it has and so on right. So, the most of these features will take only finitely many values. So, for example, If I want to do document classification. What the typical document classification problem let us, say something like Google news.

It gets feeds of lot of stories new stories and let us say it want to classify each story as entertainment or some movie story or sports story, politics whatever. So, there are so many classes given the document. I want to say the document being a a a a new story. I need to put a category to it. So, a very good feature vector is to ask for different words how many times the word occurs. So, If I get words like elections constituency Rajiv Gandhi, Rahul Gandhi then I would think it is a politics story on the other hand If I get words like balls, goals, kicks and so on. May be it is a sports story and so on.

So, as a matter of fact this word count at the feature vector is very often used in document classification which is called a bag of words representation. What we have is we choose some dictionary of words so. Me one to say twenty thousand words and then each document is represented by, a twenty thousand dimensional vector. Where each component of the vector tells you how many time the corresponding word occurs in the document. So, which means each feature then would be a discrete random variable right.

(Refer Slide Time: 31:42)



- In such cases, each feature is a discrete random variable.
- We can estimate (marginal) distribution of feature using our procedure.
- To implement Bayes classifier we need **joint** distribution of the feature vector.
- We can, e.g., assume features are independent.
- Then, joint mass function is product of marginals.
- Often called, 'naive Bayes' classifier

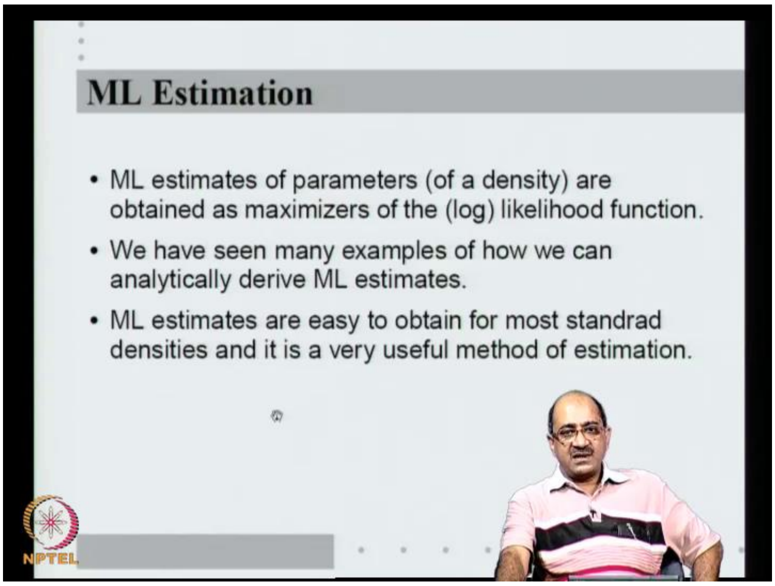
PR NPTEL course – p.76/123

In such cases, each feature is a discrete random variable the procedure. We have just now given, will tell you all such problems. We know how to estimate the mass function of individual features, because we only considered a single discrete random variables. That means, you can get marginal distribution of each of the features in all such problems. Of course to implement the Bayes classifier we need the joint distribution of the feature vector. That is the that is the class conditional density we want. So, very often what we will assume for simplicity is that features are independent. If I assume features are independent the marginal distribution tells me everything right. Then the joint mass function is simply the product of the marginal mass functions right. As a matter of fact this kind of Bayes classifier where the feature vector consists of number of features each feature is a discrete random variable.

Then I estimate the marginal distribution of each of these features right. Just by the procedure that we just now outlined and then assume that the joint mass function is simply the product of the Marginals and then use that as the class conditional density right. Here I simply assume the different the the the covariance between different features is 0.

A Bayes classifier implemented with such class conditional densities is often called the Naive Bayes classifier. The the connotation naive to say is that when, I have say in the bag of words model, I may have hundred thousand features now estimating all the Covariances is such a time consuming method, I simply neglect them. So, that is why it is called naive Bayes and in many of search and retrieval document classification and similar web based applications. A Naive Bayes classifier performs quite well as a matter of fact for the information retrieval people Naive Bayes classifier is a very important classifier. And the last example we considered in maximum likelihood estimation is how you can estimate the mass function of any of this any of such discrete features which, take only finitely many values.

(Refer Slide Time: 33:53)



ML Estimation

- ML estimates of parameters (of a density) are obtained as maximizers of the (log) likelihood function.
- We have seen many examples of how we can analytically derive ML estimates.
- ML estimates are easy to obtain for most standrad densities and it is a very useful method of estimation.

So, let us sum up what we have done about ML estimation. So, far ML estimates of parameters of a density are obtained as the maximizers of the log likelihood function log likelihood function is simply the product of the density or mass function at the values of the data. So, I take the product and then either likelihood or the log likelihood. I take I find the maximizers of the parameters and that is my m l estimate and once the m l estimate for parameters is given now class conditional density is completely known we have seen many examples, of how we can analytically derive m l estimates.

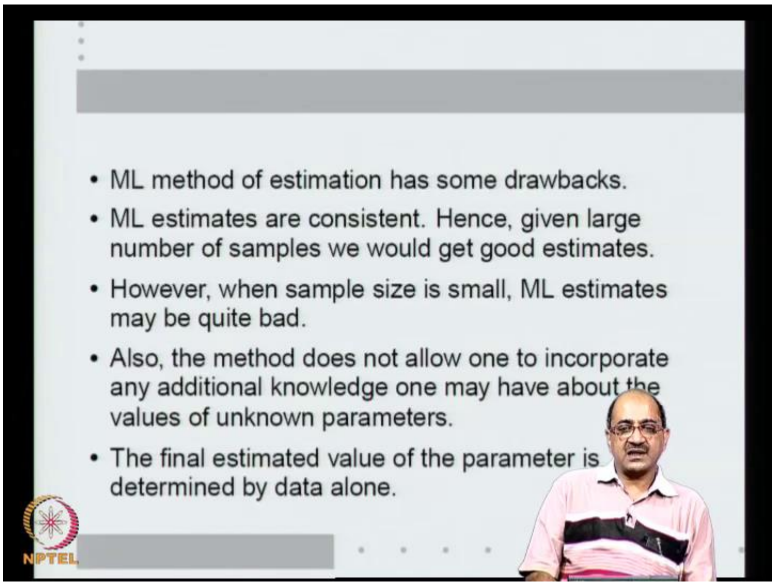
We have seen one-dimensional Gaussian density, we have seen multidimensional Gaussian density. We have seen simple example of discrete random variables. such as, Bernoulli simple examples of continuous random variables. Such as exponential right and we have also seen how in general to estimate the mass function of any discrete random variable taking finitely many values right.

So, in all these cases, we can analytically derive the ML estimates and what is also interesting is the derived ML estimates are often intuitively very satisfying see very often in any density model. The parameters are related to moments of the density all moments have some expectations and we know from law of large number. That sample means or good the I mean sample means understood in a generalized term or good approximations of expectations and that is what turns and it turns.

So, that all ML estimates versus sample means for example, the actual mean of a normal density the ML estimate for it is simply the sample mean for an exponential density. The parameter is lambda and lambda is related to the mean of the density by expected value of x is 1 by lambda and hence ML estimate for lambda transferred to be one by sample mean and so on right.

For the Bernoulli case for discrete random variable case all ML estimates are the analytically are easy to analytically derive and they always have some very simple way to understand that right. So, ML estimates are easy to obtain for the most standard densities and this is a very useful method of estimation specifically, we have discussed how using discrete random variables we can easily construct.

(Refer Slide Time: 36:25)



- ML method of estimation has some drawbacks.
- ML estimates are consistent. Hence, given large number of samples we would get good estimates.
- However, when sample size is small, ML estimates may be quite bad.
- Also, the method does not allow one to incorporate any additional knowledge one may have about the values of unknown parameters.
- The final estimated value of the parameter is determined by data alone.

So, called Naive Bayes classifier having said this ML estimation does have it is problems. Obviously, nothing is a solution for everything the reason, why we actually looked at ML estimate in the beginning is that, we said ML estimate is a very nice general procedure for obtaining consistent estimators. What are consistent estimator, estimates that converts to the true values as the sample size goes to infinity because, these are consistent, if I have large number of samples.

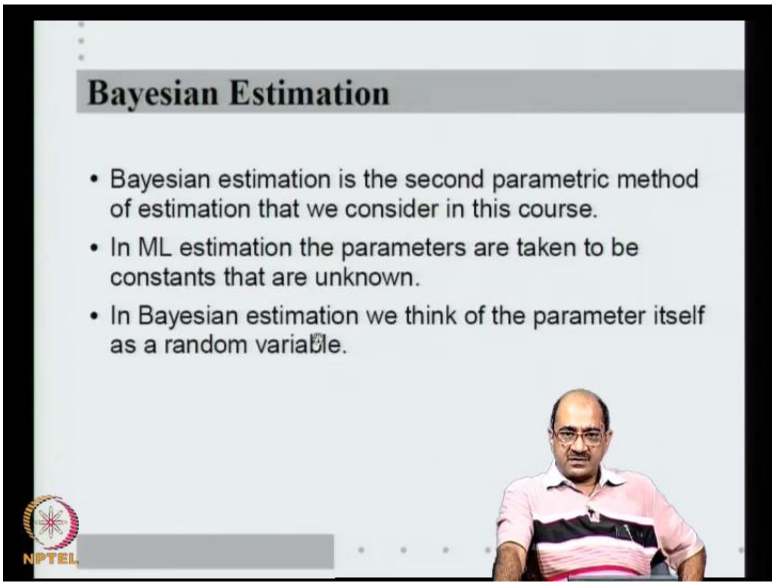
I know I will get a good estimate right. That is no problem but, what happens if I have small estimates small sample size right. The estimates can be quite bad for example, let us say I am estimating probability of heads of a coin an, I do not have too much time. So, I will toss the coin only three times and my as my luck has it all three times it turns out heads.

Then, if I want to take an m l estimate, I have to take it to be one nothing else right. That is the ML estimate for probability of heads right, because m l estimate is the sample mean we have seen the ML estimate for the Bernoulli random variable. Now that is obviously, not very good I know that the coin is not not bias to that extent. But I can do nothing because for small sample size that is how Ml estimate turn out to be what it means is even.

If I have some knowledge about the possible values for the parameter in this case I may know that the coin is not a two headed coin both probability of heads and tails are strictly greater than 0. But that of course I do not know the actual probability of heads but, I know that the probability of heads as well as, the probability of tails has to be strictly greater than 0. But in within the Ml estimate I have no way of incorporating this extra knowledge that I have right.

That is because, the final estimated value is simply determined by the data it determined by the data alone and nothing else. If I got three, I toss the coin three times and I got three heads the maximum likelihood procedure does not leave me any other option. But, to say probability of heads is equal to one. There is now way I can tell the estimation procedure that, I know that the probability of consistency is less than 1 right. There is another method of estimation, which we are going to consider of course, there are many other methods. But we have considered only two methods.

(Refer Slide Time: 38:49)



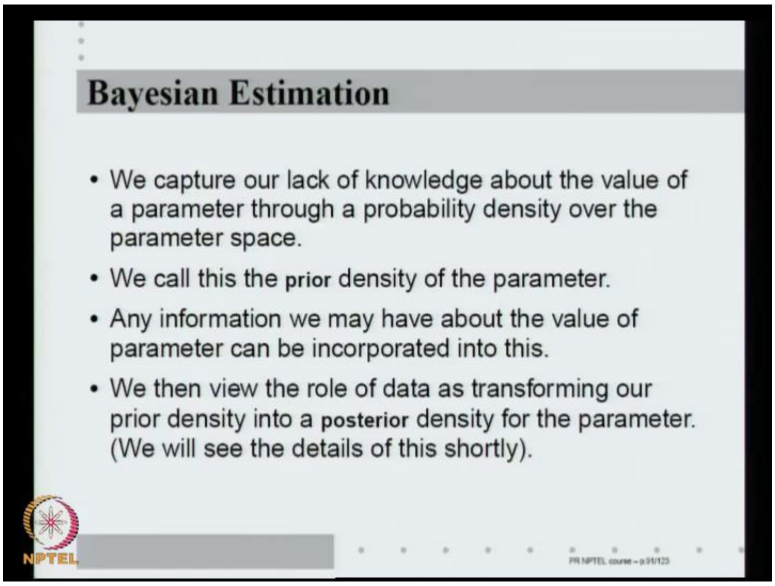
Bayesian Estimation

- Bayesian estimation is the second parametric method of estimation that we consider in this course.
- In ML estimation the parameters are taken to be constants that are unknown.
- In Bayesian estimation we think of the parameter itself as a random variable.

So, the second method of estimation that we will consider in this course is called the Bayesian estimation. It is also a parametric method to some extent, this actually takes care of some of these problems with ML estimate. So, that at for small sample sizes this actually gives you better estimates. At the top level the way, I can distinguish between a maximum likelihood and Bayesian estimation that is in the maximum likelihood estimation.

The parameters are taken to be constants that are unknown right; while I do not know the value of the parameter theta. Theta is not random; theta is taken to be a constant in the Bayesian estimation. We think of the parameter itself as a random variable. This is the basic difference between the Bayesian and maximum likelihood estimation, we think of the parameter itself as a random variable.

(Refer Slide Time: 39:49)



Bayesian Estimation

- We capture our lack of knowledge about the value of a parameter through a probability density over the parameter space.
- We call this the **prior** density of the parameter.
- Any information we may have about the value of parameter can be incorporated into this.
- We then view the role of data as transforming our prior density into a **posterior** density for the parameter. (We will see the details of this shortly).

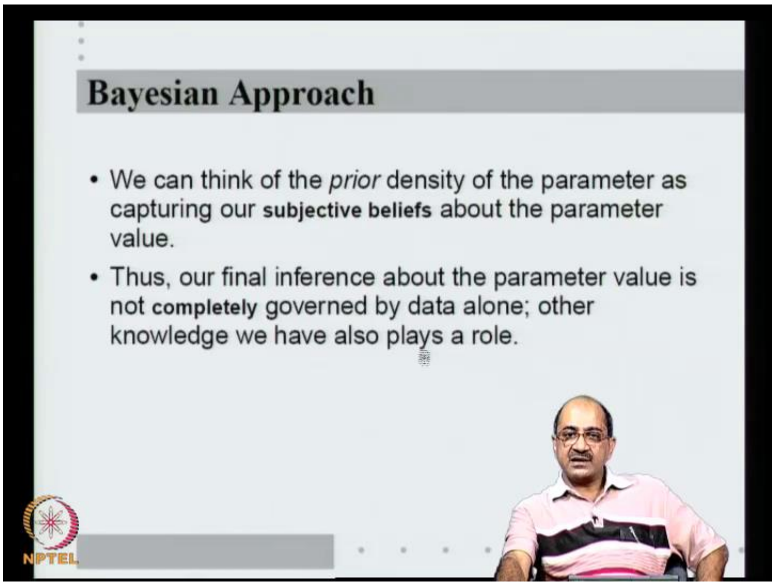
PR NPTEL course – p.31/123

Then because, we think of the parameters are random variable, we say that our lack see we we are doing estimation because, we do not know the knowledge of the parameter value. So, we capture our lack of knowledge about the value of parameters that is the fact that the parameter can take different values, because we do not know what specific value. It takes even though, we know the range of values it can take right. So, we capture our lack of knowledge about the value of the parameter through, a probability density over the parameter space or all possible value. That the parameter can take we put a probability density function and we want this density function to capture what our knowledge or lack of it that, we have about the possible values that the parameter can take such a density is called prior density of the parameter.

Prior in the sense is before, I saw any data before seeing any data. I have some model some density model for the parameter. Parameter is assumed as, a random variable and this is my the prior density is the density of the parameter. Before, I saw any data any information that we have about the value of the parameter can be incorporated to this density model right.

Then we view the role of data as transforming our prior density; this is our view of what the parameter could, take into a posterior density; posterior meaning after seeing the data right. So, we think of the role of data in estimation as transforming a prior density into a posterior density for parameters. We will see many examples of that in the next class by shortly, I mean may be next class but, the rest of this class, we will look at just this view point in a little more detail.

(Refer Slide Time: 41:39)



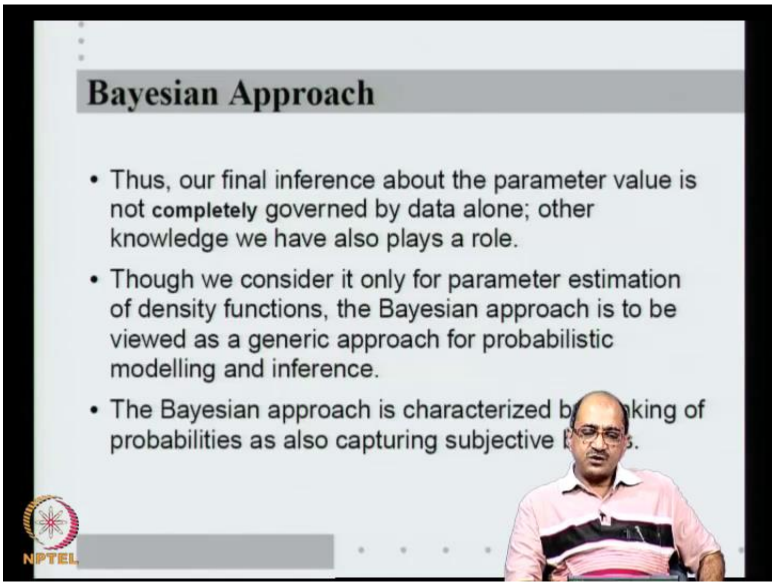
Bayesian Approach

- We can think of the *prior* density of the parameter as capturing our **subjective beliefs** about the parameter value.
- Thus, our final inference about the parameter value is not **completely** governed by data alone; other knowledge we have also plays a role.

We think of the prior density of the parameter as capturing, what can be called subjective beliefs about the parameter value. The Bayesian approach is an approach to probability not just about parameters. The approach is that one can capture through probabilities is our subjective beliefs right. That is why we were thinking of prior density of parameter as capturing our subjective beliefs about the parameter value at this level.

I will i will explain a little more presently but, what it would mean is our final inference about the parameter value is not completely governed by data alone as an m l estimate the prior subjective beliefs. I have about the parameter also play a role in making out final inference about the parameter value given the data. So, a final inference is not completely governed by, the data alone other knowledge we have also plays a role this view is essentially an approach.

(Refer Slide Time: 42:47)



Bayesian Approach

- Thus, our final inference about the parameter value is not completely governed by data alone; other knowledge we have also plays a role.
- Though we consider it only for parameter estimation of density functions, the Bayesian approach is to be viewed as a generic approach for probabilistic modelling and inference.
- The Bayesian approach is characterized by thinking of probabilities as also capturing subjective beliefs.

So, let me put this again our final inference is not completely governed by data alone. Other knowledge also plays a role. This is an approach not just though we considered only for parameter estimation. This is an approach to probability the approach is characterized by, thinking probabilities, as also capturing subjective beliefs. We can broadly distinguish between, what can be called a Frequentist approach probability versus.

What it called a Bayesian approach to probability the Frequentist approach we think of probability as a limiting value or the fraction of times an event occurs. If I repeat the random experiment there are lot of things, which we loosely talk in probabilities terms, which are not really that kind of repeatable random experiments. We talk of probability that it rains tomorrow because, tomorrow. But comes only once; I cannot in principal also repeat the experiment enough times.

So, whatever probability, I talk about probability of rain tomorrow is more like a subjective belief of course, might be based on know some data and some sound scientific principles but, it still something that captures my belief or my ignorance about whether or not it rains tomorrow. So, the idea is that probabilities can also capture lack of knowledge through, whatever you know subjective inference that we can make apart from only capturing situations of repeated random experiments.

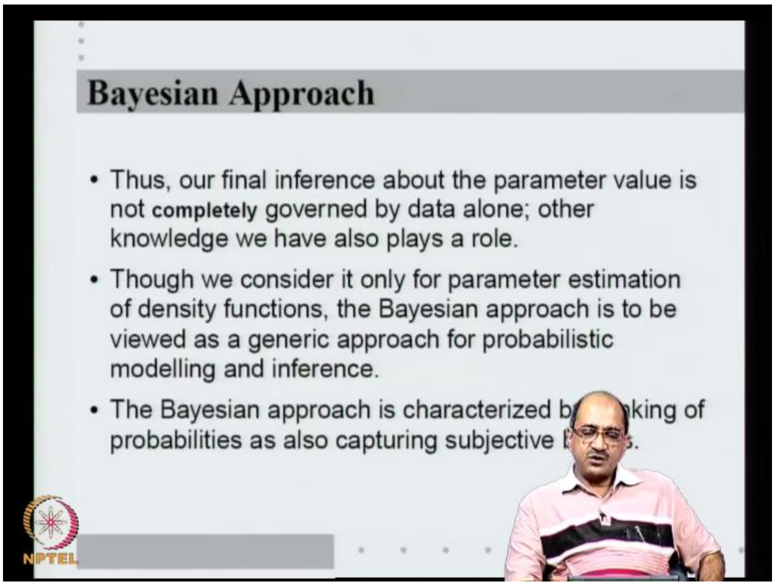
Yeah may be at this stage, I can give you to a classical example of Bayesian approach among this of course, given by a Bayesian statistician. So, it is heavily tilted for the Bayesian approach but, still is an instructive example, consider three situations of time to make inference based on gathering data based on making an experiment. The first situation is there is a musician, who says if you just show me the notes not the words not the full score but, just a line of notes of a of a song. I can tell you whether or not, it is composed by some famous musician, let us say whether or not, it is composed by Tyagarajan. So, we do some test. So, we just write one line of notes give it to him. So, let us say in a experiment conducted three out of three. He correctly guessed, whether or not what is given to him is composed by Tyagarajan.

Now, situation two there is a lady who says, if she tastes, if she tastes tea. She can tell you whether, you pour decoction into cup containing milk or milk into a cup containing decoction. Once again you conduct an experiment three out of three times, she guessed correctly my third situation. There is a man a drunk, who says when he is fully drunk, he can correctly predict whether a fair coin comes heads or tails you conduct an experiment let us assume. Once again three out of three times he he he guessed correctly now, what is the inference, if I go by data alone like in the maximum likelihood.

I have to either grant all three claims or not grant any of them saying data is insufficient. But, what most of us tend to do is that may be grant the first claim dismiss away the last claim as a lucky, a lucky, a lucky run and shrug our shoulders, about the middle right; which is simply based on our subjective belief of what is possible and what is not possible. Of course, I had to think to make it interesting, let us say the experiment is five out of five times. They guess correctly, even then we will certainly, when it becomes five out of five we were much more confident in granting the claim of the musician.

We will still dismiss away the last claim as a lucky as a lucky run after all, you know there have been many cricket series, where Captain always wins all the tosses. So, they can be lucky runs of five where you call, call a coin correctly and we may still shrug our shoulders about the lady really had that ability or not. So, the idea is that data can speak only so much right. There is also our subjective prior belief on the situation, which modulates our inference from the data of course, in this particular example. If I increase the data to hundred then, I have to grant my third person is extra sensitive perception if it indeed. So, happen that he can correctly guess right. So, the idea is that this our subjective belief can modulate what the data is trying to speak to us now this is the basic Bayesian approach right.

(Refer Slide Time: 48:08)

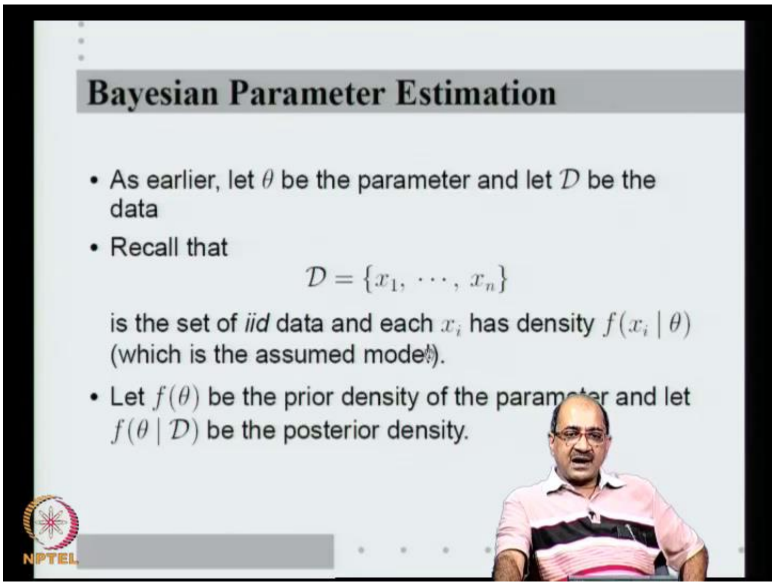


Bayesian Approach

- Thus, our final inference about the parameter value is not **completely** governed by data alone; other knowledge we have also plays a role.
- Though we consider it only for parameter estimation of density functions, the Bayesian approach is to be viewed as a generic approach for probabilistic modelling and inference.
- The Bayesian approach is characterized b[illegible]king of probabilities as also capturing subjective [illegible]s.

So, we have So, in terms of the estimation of parameters, what it means is I have a prior subjective density. Over the parameters and the data transforms my prior density into a posterior density. So, coming back this will make sure that our final inference is not blindly only about data but it also depends on the prior density.

(Refer Slide Time: 48:16)



Bayesian Parameter Estimation

- As earlier, let θ be the parameter and let $\mathcal{D}$ be the data
- Recall that

$$\mathcal{D} = \{x_1, \cdots, x_n\}$$

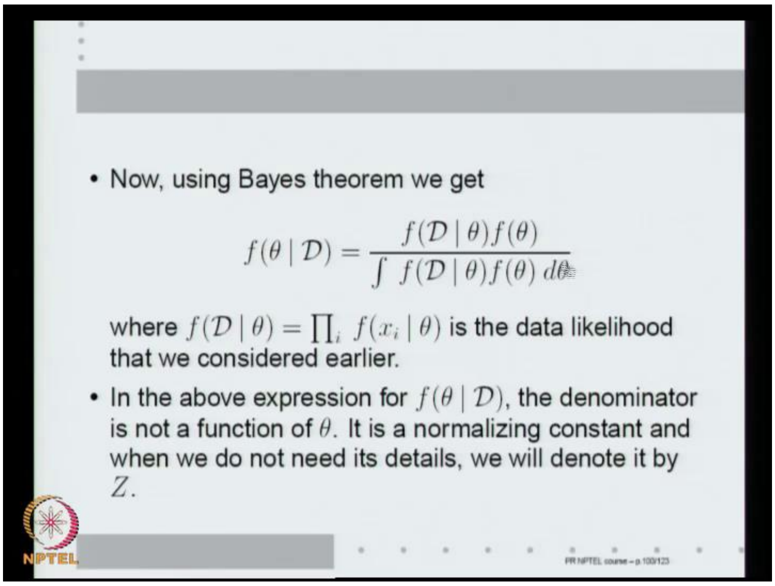
is the set of *iid* data and each x_i has density $f(x_i \mid \theta)$ (which is the assumed model).

- Let $f(\theta)$ be the prior density of the parameter and let $f(\theta \mid \mathcal{D})$ be the posterior density.

So, let us leave all the philosophy behind and come back to only the parameter estimation problem. So, as earlier let theta be the parameter that we need to estimate and skip D be the data. Data as usual is n iid realizations x 1 to x n the set of iid are and each x i has density f x i given theta, f x i given theta is the assumed density model. Bayesian estimation is also a parametric estimation.

So, there is an assumed density model f x i given theta. So, x 1, x 2, x n are the datas. So, what we have extra now is there is a prior density and theta. So, f theta is a density over the random variable theta. So, our theta is viewed as a random variable, which has a prior density we represented by f and the posterior density is the density of theta conditioned on d after seeing d my density becomes theta.

(Refer Slide Time: 49:14)



- Now, using Bayes theorem we get

$$f(\theta \mid \mathcal{D}) = \frac{f(\mathcal{D} \mid \theta) f(\theta)}{\int f(\mathcal{D} \mid \theta) f(\theta)\, d\theta}$$

where $f(\mathcal{D} \mid \theta) = \prod_i f(x_i \mid \theta)$ is the data likelihood that we considered earlier.

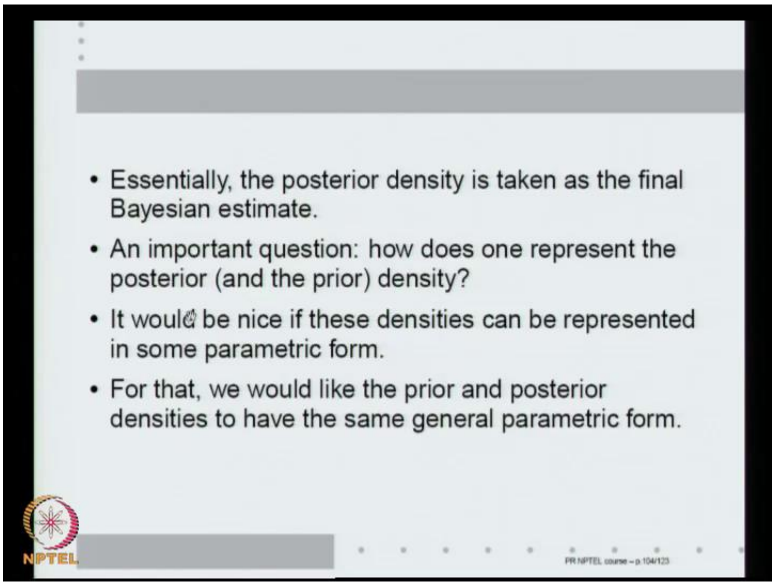
- In the above expression for $f(\theta \mid \mathcal{D})$, the denominator is not a function of θ. It is a normalizing constant and when we do not need its details, we will denote it by Z.

So, now how do I get the posterior density we use Bayes rule; so using Bayes rule. I am assuming everybody knows Bayes rule for density functions just like Bayes rule for probabilities. So, f of theta given D is given by, f D given theta into f theta by normalizing factor integral f D given theta f theta D theta the theta in the denominator ofcourse, is a dummy variable because, I am integrating with respect to theta, I could as well have called the theta prime. So, the denominator is a constant. So, the posterior density f theta given d is propositional to the product of f d given theta into f theta this is the Bayes rule. What is f d given theta because, D is x1, x2, x n and their iid it is simply f of x i given theta this is the data likelihood right.

So, the posterior density is given by, the product of the prior and the likelihood function. As I said in the expression, the denominator is not a function of theta. So, we remember that is only a normalizing constant. So, whenever we do not need the details of expression we simply replace it by some constant let us say capital Z. So, what happens is even in Bayesian likelihood, we will calculate the likelihood function there, we were simply asking which theta maximizes this instead of that we are obtaining our prior density.

At that density, which is proportional to the product of the likelihood our obtaining our posterior density at the density, which is propositional to the product of the likelihood and the Prior density. Of course to make it a density, we need a normalizing constant that is given by this. So, we essentially given data on our model, because I know f x i given theta. I know how to calculate f D given theta, I am given a prior density. So, I take the product and turn it into the density that is my posterior density.

(Refer Slide Time: 51:10)



- Essentially, the posterior density is taken as the final Bayesian estimate.
- An important question: how does one represent the posterior (and the prior) density?
- It would be nice if these densities can be represented in some parametric form.
- For that, we would like the prior and posterior densities to have the same general parametric form.

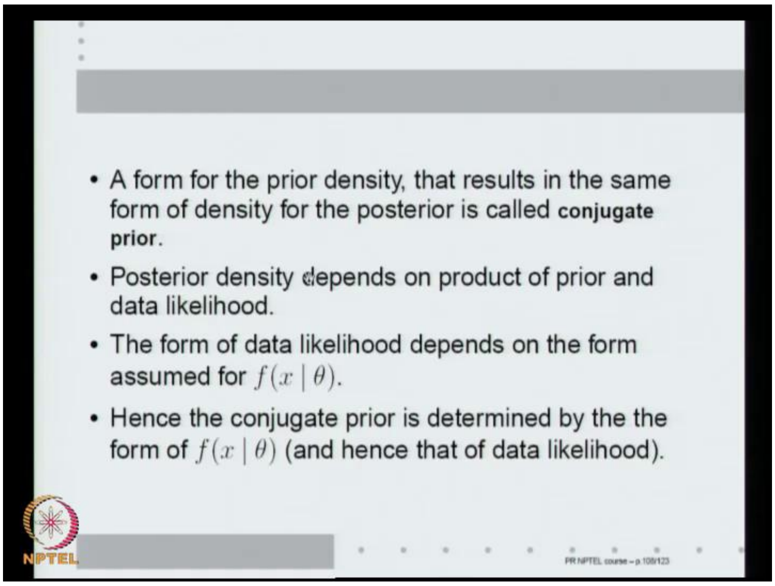
PR NPTEL course – p.104/123

So, in the Bayesian estimation my posterior density is the final Bayesian estimate. We will see of course, how do I find the posterior density and you know there are many other issues with it. But today's class we will just look at you know the broad outlines. So, we take the posterior density as, the final estimate then come to the problem how do I calculate this. If this is some density over theta, and I multiply this; so for every given value of theta may be I can calculate this. But I cannot represent it as an infinite table because, theta is a continues value right.

So, there is a problem of how do I represent the prior and posterior density right. So, the important question because, we want to take posterior density as the final estimate is how does one represent the prior and posterior densities. It would be nice if these can be represented in terms of parametric form. So, then we will simply store the corresponding parameters values for these densities. For that what we would like is that the prior and posterior densities should belong to this same parametric formulization. The prior is say

Gaussian density, we would like the posterior also to be a Gaussian density. Otherwise, every time, I get more data, I may get a different density.

(Refer Slide Time: 52:19)

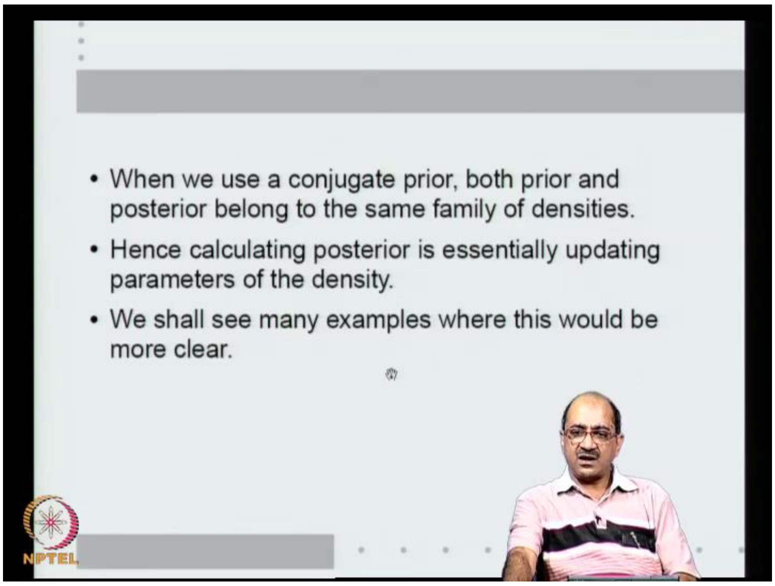


- A form for the prior density, that results in the same form of density for the posterior is called **conjugate prior**.
- Posterior density depends on product of prior and data likelihood.
- The form of data likelihood depends on the form assumed for $f(x \mid \theta)$.
- Hence the conjugate prior is determined by the the form of $f(x \mid \theta)$ (and hence that of data likelihood).

So, this means I should choose my prior carefully a form of the prior density. That results in the same form of density for the posterior is called conjugate prior right. The conjugate prior is a density for the prior; that result in the same form of density for the posterior. Now the posterior depends on the product of prior on the likelihood and the form of likelihood depends on the form of the assumed model f x given theta.

So, ultimately, what is a conjugate prior depends is determined by the assumed model of the data likelihood. So, for a given model and hence a given data likelihood a different for a for a given thing. There will be some given density for p theta that becomes conjugate; so when we do Bayesian estimation, we will chose the right prior. So, that it is a conjugate prior, we will see many examples of this next class.

(Refer Slide Time: 53:11)

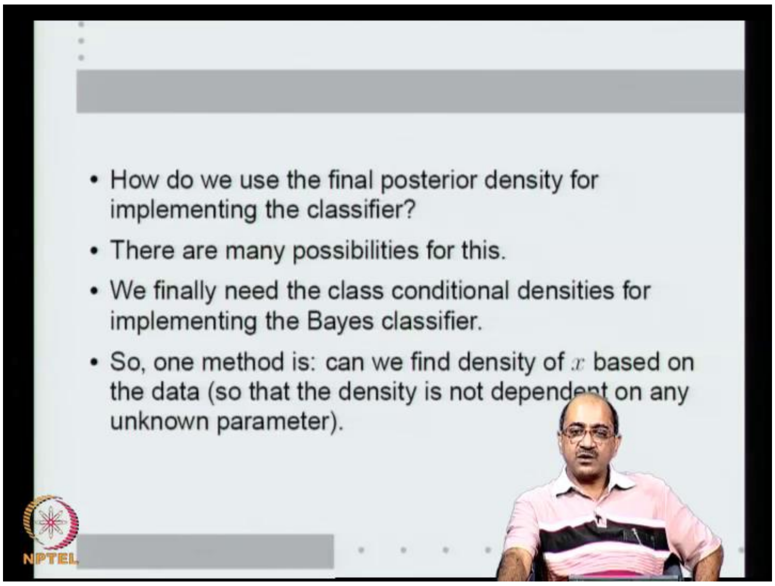


- When we use a conjugate prior, both prior and posterior belong to the same family of densities.
- Hence calculating posterior is essentially updating parameters of the density.
- We shall see many examples where this would be more clear.

But the main issue here is when we use a conjugate prior. Both prior and posterior belong to the same family of densities. So, prior might be some normal with mean mu 0 on variance sigma0 and posterior will be, also normal only thing that may change is mean and variance. So, may be mean sigma n and mean mu n and variance sigma n. So, then calculating the posterior is essentially updating the parameter density given the data. And hence the data likelihood and the prior density meaning give me mu 0 on sigma 0. Then the calculation simply involves, how to calculate mu n and sigma n namely the parameters of the posterior density, because we know prior and posterior will be of the same family.

This is what, happens when you use conjugate prior and that is the reason, why when we do Bayesian parameter estimation. We will always assume conjugate prior but, conjugate prior is not a single density for a particular model there is, when I am estimating mean of a Gaussian something may be conjugate prior. When I am estimating a Bernoulli density parameter of a Bernoulli density some other density may be a conjugate prior and so on. So, for each case we have to find out what the right. Conjugate prior is we shall see many examples on how this is done that is, when it becomes more clear.

(Refer Slide Time: 54:27)

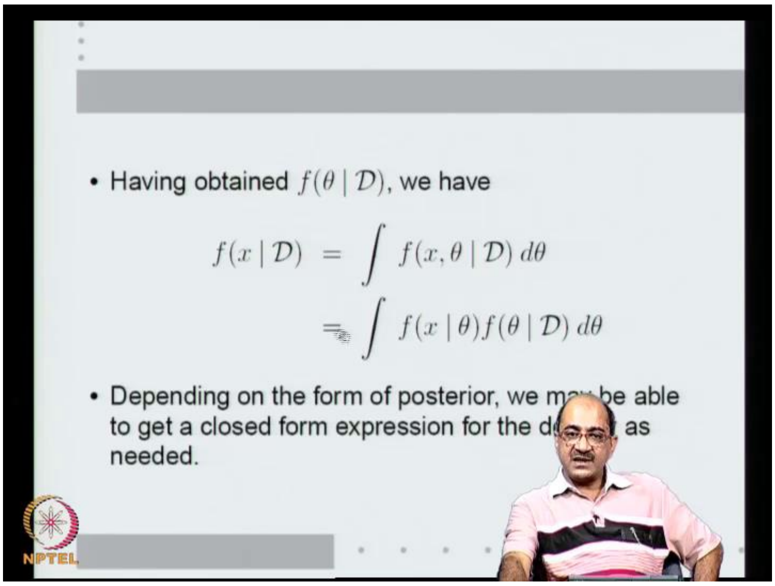


- How do we use the final posterior density for implementing the classifier?
- There are many possibilities for this.
- We finally need the class conditional densities for implementing the Bayes classifier.
- So, one method is: can we find density of x based on the data (so that the density is not dependent on any unknown parameter).

Now, you want to take the posterior as your final answer that is the, that is the end of the estimation. So, what does that mean if I have the posterior density. Over theta how can I use it in a classifier. My classifier actually needs, class conditional densities right. Ultimately somebody has to give me class conditional densities. So, that is the reason m l estimate gives me a particular value of a theta.

I plug in that that theta in my model that is my final class conditional density. But, once we will think of parameter as a random variable and our posterior densities. There is more than a one possibility for this. One method is can I use the posterior. So, that I finally, get some density of x based on the data, which is not dependent on any other unknown parameter this is possible.

(Refer Slide Time: 55:14)



- Having obtained $f(\theta \mid \mathcal{D})$, we have

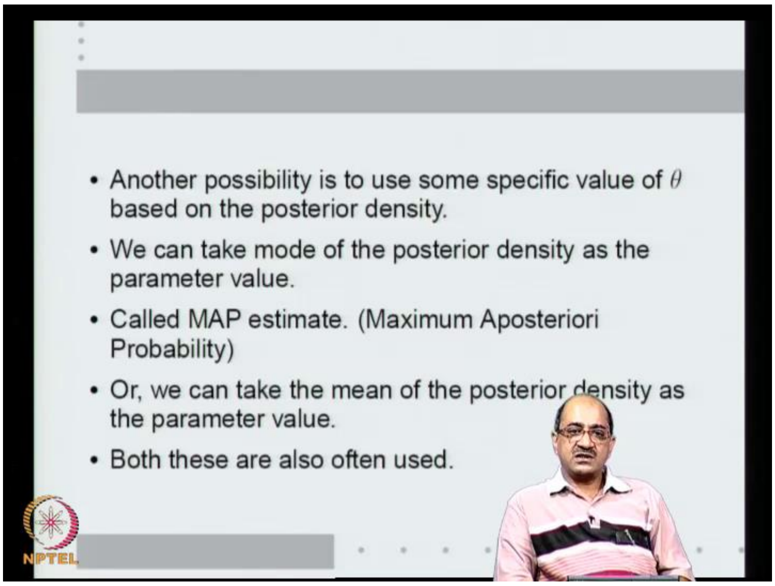
$$f(x \mid \mathcal{D}) = \int f(x, \theta \mid \mathcal{D})\, d\theta$$
$$= \int f(x \mid \theta) f(\theta \mid \mathcal{D})\, d\theta$$

- Depending on the form of posterior, we m be able to get a closed form expression for the d as needed.

Once I have the posterior density f theta given d. Now I can write the density of x conditioned on D as density of x, and joint density of x, and theta conditioned on D integrated with respect to that this is that. I am obtaining the density of x by marginalizing or theta of the joint density of x, and theta. Now I can write the joint density at the product of conditional, I can write this as a x given f of x given theta D into f of theta given D and x. Given theta D is same as x given theta because.

I know once given theta there data, does not is not needed to find the density of x. The only thing unknown is a density of x is theta. So, f x given theta d becomes x given theta. So, this integral becomes integral f x given the f of x given theta into f of theta given d d theta. So, if I know the posterior, I know this model. I can integrate with respect to theta. I get something that is only dependent on data right theta is now gone. So, this is what I can use actually as, a class conditional density based on the data D of course. This may not always be possible, only if this model is nice and if this prior density is nice. We may be able to this integral to get this into a nice form that, can be implemented when when it is possible there is one way of using, the Bayesian estimation for getting class conditional densities.

(Refer Slide Time: 56:36)



- Another possibility is to use some specific value of θ based on the posterior density.
- We can take mode of the posterior density as the parameter value.
- Called MAP estimate. (Maximum Aposteriori Probability)
- Or, we can take the mean of the posterior density as the parameter value.
- Both these are also often used.

Otherwise, we simply have to take some specific value of theta based on the posterior density for that also there are multiple possibilities. One is we can take the mode of the posterior density that is the value of theta at which p theta, given D is maximum right. The posterior density is maximum the value of theta that maximize the posterior density is one value. That I can take this is often called the map estimate maximum aposteriori probability estimate.

The map estimate is also something that is very specific. Only to Bayesian estimation, because it depends on the mode of the posterior density; so it is actually the mode of the posterior density or we can simply take the mean of the posterior density as the parameter value this is also possible; so essentially looking at the posterior density. We take some specific value of theta obtained from the posterior density and add the final value of theta plug. That in and get our class conditional densities all these are used. So, this is the general introduction to the Bayesian estimation. So, next class we will look at a few examples of how this whole thing is done how in specific situations.

We get conjugate priors, how we may be able to find f of x, given D how we may be able to find map or expectation of the posterior at the estimates. We will once again do for the same densities for normal density for Bernoulli for general finite value discrete random variables and so on. So, that we can compare the what the answers we get with Ml estimation with the answers, we get with Bayesian estimation.

Thank you.